Rijksuniversiteit Groningen

Bachelor Thesis

Investigating stellar streams in the retrograde Milky Way halo

Selection of Phlegethon members with a reduced proper motion selected halo catalogue as a proof of concept

rijksuniversiteit groningen

Author:
Anna. F. Esselink

Supervisors:
dr. Else Starkenburg
Akshara Viswanathan

Abstract

This thesis aims to detect stellar streams in the retrograde Milky Way halo using the Reduced Proper Motion selected halo sample created by Viswanathan et al. (2022) using data from Gaia EDR3. The sample provides photometric distances, that are more reliable than the Gaia parallaxes, for around 48 million stars. For seven globular clusters that were identified in the sample, the mean photometric and metallicity-dependent distance were computed. This provided reliable distances for three globular clusters. For the other clusters, the stars in the sample belonged to the main sequence turn-off instead of the main sequence. Consequently, their magnitudes were incorrectly assigned, resulting in incorrect distances. The color cut $0.45 > G - G_{RP} > 0.715$ was shown to effectively remove the majority of stars corresponding to the main sequence turn-off. To systematically find stellar streams the sample was probed for different spatial and proper motion intervals in the outer halo to create a series of plots. The streams GD-1, Jhelum, Indus, Ophiuchus, Ylgr, and Orphan were identified through these plots. However, more conditions need to be in place to efficiently find and identify streams with this method. Candidate members of faint, strongly retrograde stream Phlegethon, located in the solar vicinity, were selected using a series of alternating polynomial fits to overdensities in proper motion and the stream-track. 575 stars were selected with a mean photometric distance of 3.6 ± 0.6 kpc. Using a crossmatch with Pristine, a metallicity of $[\mathrm{Fe/H}] = -1.9 \pm 0.5$ was established. The stream probes magnitudes up to $m_G \approx 21$, which is a magnitude fainter than previous selections of Phlegethon by R. Ibata using the STREAMFINDER algorithm. By not constricting stars to an orbit, more features of the stream, like spurs are likely to be observed. The selection of Phlegethon candidates shows that stellar streams can be selected through a method that can be automated. This will eventually help to create an algorithm that can systematically and automatically detect stellar streams with the 5D data sample.

Contents

1 Introduction

Galactic Astronomy is the study of the Milky Way and its contents. By studying our Galaxy we can gain insight into its structure and history. Due to Milky Way being a fairly average spiral galaxy (Bland-Hawthorn & Gerhard, 2016), it can also be used to study other systems. It can provide a better understanding of galaxy formation and even of the formation of large-scale structures in the Universe.

1.1 Components of the Milky Way

The main components of the Milky Way are the thin disk, the thick disk, the bulge/bar, and the stellar halo, shown in fig. 1. Each component has its own distinct characteristics.

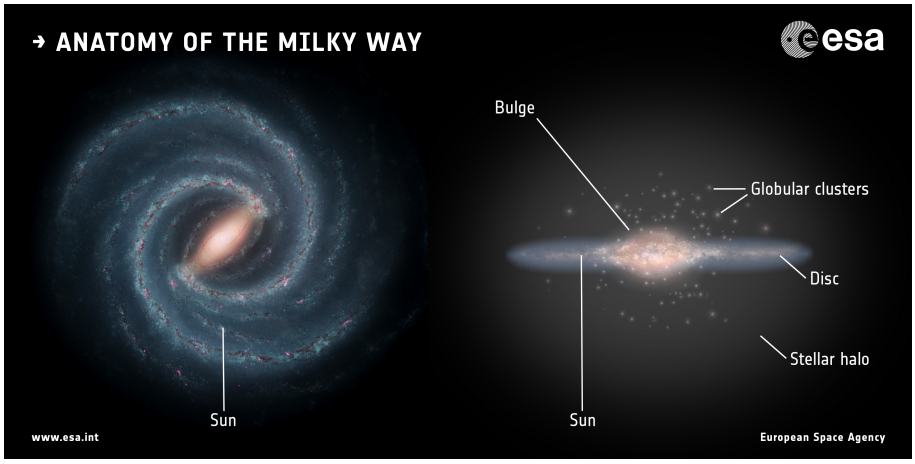


Figure 1: An artist's impression of the Milky showing its main components. Image credits: Left: NASA/JPL-Caltech; right: ESA; layout: ESA/ATG medialab

The thin disk is the most noticeable feature of the Galaxy. There is active star formation at an estimated rate of $\sim 1.6\,\mathrm{M}_\odot/\mathrm{yr}$. It contains mostly young stars that move in approximately circular orbits. In addition, it extends up to 16 kpc and has a scale height of ~ 300 pc. The thick disk, as the name suggests, is thicker than the thin disk and has a scale height of ~ 1 kpc. It is also more diffuse and hotter, and contains older stars than the thin disk Bland-Hawthorn & Gerhard (2016). Stars from the thin and thick disk can be separated by their chemical components, specifically [α/Fe] and [Fe/H]. α-elements like O, Mg, Si, Ca, S, and Ti are released during supernovae of massive stars (type II). Fe is produced in supernovae of a white dwarf in a binary system (type Ia). Supernovae of type Ia happen over a longer timescale than type II, due to the lower mass of the former. Hence [α/Fe] is expected to decrease over time. The thin disk corresponds to low α values and the thick disk to higher α abundances (Bensby et al., 2003).

The bar/bulge is the densest component, located in the center of the Galaxy. It is obscured by gas and dust, which interferes with observations and limits our understanding of this component. The bulge contains around 40% of the total stellar mass in the Galaxy (Valenti et al., 2016). In the true center resides a supermassive black, namely Sagittarius A*. For a long time, it was unclear whether the Milky Way had a bar. Its existence has now been confirmed, however, there are still debates about its properties, like the size and rotation (Wegg et al., 2015; Portail et al., 2017). The bulge contains different stellar populations, including very old (> 13 Gyr), and metal-rich populations. Other populations correspond to the other Milky Way components.

The Milky Way halo is the most extended region of the Galaxy and is spherical in shape. The halo is dominated by dark matter contained in the dark halo. The stellar halo makes up only $\sim 0.1\%$ of the halo's mass Girelli et al. (2020). In this thesis, the focus will be on the stellar part. The stars in the halo are generally old and metal-poor. The halo plays an important role in the study of the history of the Galaxy, known as Galactic Archaeology. For example, tidal streams provide information about their progenitors, globular clusters, and dwarf galaxies that have merged with the Milky Way.

1.2 Galaxy formation and evolution

The currently accepted cosmological framework is the ΛCDM model in which structure is formed hierarchically inside dark matter halos (White & Rees, 1978). Proto-galaxies grow through a series of mergers with other systems to form the galaxies that are observed now. Star formation in galaxies is fueled by gas coming in from mergers and the intergalactic medium. In galaxies like the Milky Way, roughly half the baryonic mass originates from mergers and the other half from the intergalactic medium. However, mergers are responsible for bringing in the majority of the dark matter (Wang et al., 2011) and only $\sim 10\%$ of the stars in the Galaxy. This suggests gas-rich mergers and a subsequent higher star formation rate (Rodriguez-Gomez et al., 2016).

The formation and dynamics of galaxies are strongly dependent on their dark matter halos. The halo attracts baryonic matter, in the right conditions the gas will cool down and begin star formation. As the gas collapses towards the center of the halo, a rotating gaseous disk can form as consequence of angular momentum conservation (Mo et al., 1998).
In the past, mergers happened more frequently, due to the higher density of the Universe. Large merger events could also contribute to the creation of a bulge (Barnes, 1992). In the case of the Milky Way, it led to the formation of the stellar halo consisting of stars from the original disk and the progenitor, named Gaia-Enceladus (Helmi et al., 2018). It is likely that mergers also disrupted the formation of the thin disk. Only after the activity decreased was it able to grow further. Consequently, stars in the thin disk are commonly younger than in the stellar halo.
A large merger event could also be responsible for the creation of a bar by causing instability in the disk of the Galaxy. The exact origin of the bar is unknown. The bar could have completely originated from the thin disk or for some part also from the primordial thick disk Martinez-Valpuesta & Gerhard (2013).

These examples show that the formation history of the different galactic components is connected. A large merger event could be responsible for multiple structures. Based on the properties of the stars in the different components, one can extract information about their origin. By determining the sequence of events that occurred, the formation of the Galaxy can be revealed Helmi (2020).

1.3 Galactic Archaeology

Identifying stellar streams and their properties reveals information about the merger history of the Milky Way. Based on their position and orientation, an estimate of their orbit can be made, based on which an estimate of the distribution of their progenitors can be made (Grillmair & Carlin, 2016). Streams are also important in the hierarchical cosmological paradigm. Mergers form the dynamical mass (the dark halo) of a galaxy. They also provide information about the general build-up of galactic systems.

The luminous satellites that are accreted in the Galaxy can generally be associated with one of two types: dwarf galaxies and globular clusters. Dwarf galaxies have similar features as regular galaxies but are smaller in size. Over 30 dwarf galaxies have been found in the Milky Way (McConnachie, 2012). The remnants of dwarf galaxies in the Milky Way and other observed dwarf galaxies have a similar size and mass, however, their chemical composition is different (Tolstoy et al., 2003). This can be explained by the fact that observed dwarf galaxies had Gyrs longer to evolve and form stars than the accreted dwarf galaxies. This theory is supported by current cosmological simulations (Fattahi et al., 2020).
Globular clusters (GCs) are dense stellar systems containing stars with similar ages and metallicities. Over 150 GCs have been found in the galactic halo (Harris, 1996). The GCs can be split into two groups based on their origin, the ones that formed in the Milky Way and the ones accreted. The majority of GCs are thought to originate from accreting dwarf galaxies (Renaud et al., 2017).

Chemical abundances can be used to distinguish different systems. Based on the environment and timescale in which the stars were formed, each system has its own distinct chemical sequence. This is observed for dwarf galaxies in the Local Group. The same principle applies to dwarf galaxies that were accreted. Based on the mass and the time galaxies had to form stars, their [α/Fe] and [Fe/H] abundances will differ. E.g. high [α/Fe] at low [Fe/H] will correspond to low mass galaxies with a single generation of stars, while galaxies with star formation on a large timescale could have high [α/Fe] at low [Fe/H] (Tolstoy et al., 2009). Chemical abundance analysis like this can be used to distinguish large accreted systems. However, to identify smaller systems like accreted globular clusters, extremely accurate abundances of many elements for a large sample of stars would be needed (De Silva et al., 2006). In the case of stellar streams, it is therefore often easier to identify them based on their dynamics.

When systems like dwarf galaxies and globular clusters merge with the Milky Way, they can create a stellar stream due to the object being broken up by tidal forces under the gravitational potential of the Milky Way. The stars contain information about their past through their kinematic and chemical properties (Johnston, 1998). Stars originating from the same dwarf galaxy or globular cluster will have similar orbits due to their originally similar positions and velocities. If the original system is small or the stream has recently formed, the stream is most likely to be long and narrow (Johnston, 2016). Larger systems can lead to more complex and broad streams that are more difficult to identify, due to the larger velocity dispersion and range of energies amongst these stars (Quinn, 1984).
Identifying stellar streams helps to constrain the Galactic potential and density distribution of the halo using their dynamics. One method that is used for this, fits the orbits of streams using 6D data (Nibauer et al., 2022). Accurately measuring the gravitational field, will also allow to test the validity of the ΛCDM model.

A stellar stream is dependent on three major components, the size of the original system R and its velocity dispersion σ, the time since the formation started t, and the characteristic orbital timescale t_{orb}. An estimate of the density of a stream ρ can be given by $\rho \propto (t_{orb}/t)^3/(R\sigma^2)$. Initially when $t \sim t_{orb}$, the density will be high. Consequently, these streams will be easier to detect due to their over-densities. This property applies to many streams in the outer halo due to their long orbital time scale. However, for streams in the inner halo or streams originating from a smaller system, the density will decrease quicker due to their shorter orbital time scale and size respectively. Consequently, many accreted systems in the inner halo are very phase-mixed (Helmi & White, 1999). Often multiple streams originate from the same progenitor. In this case,

it is more difficult to find streams through over-densities in spatial coordinates. Instead, energy and angular momentum are used under the assumption that stars with the same progenitor have similar integrals of motion. However, in this thesis, the focus will be on stellar streams that are spatially coherent.

When looking for tidal streams in the vicinity of the Sun, it is estimated that the minimum data sample needed to detect at least 10 stars per stream, contains 5000 halo stars with a velocity resolution of ~ 13 km/s. This is based on the velocity dispersion of the halo (~ 100 kms/s) and on the assumption that if the halo has formed through mergers, ca. 500 streams would exist in this region. These conditions were almost met by Gaia DR2 (Gómez et al., 2013). For a more detailed characterization of the streams and their progenitors, more data with higher precision is needed. The recently release Gaia DR3 sample improves upon Gaia DR2 by giving line-of-sight velocities for 33 million stars with G_{RVS} magnitudes up to 14 (Gaia Collaboration et al., 2022).

1.4 The Milky Way halo and its assembly history

The stellar halo contains a lot of information about how the early Milky Way was formed. By tracking stellar streams in the halo a lot can be revealed about the merger history of the Milky Way. The stellar halo is also interesting due to its old and metal-poor stars. This suggests that accreted dwarf galaxies were more metal-poor than the proto-Milky Way due to being less massive (Tremonti et al., 2004). Due to the halo containing more old stars in proportion to the other galactic components, it also gives valuable insights about the physical conditions of the early universe.

With data from Gaia DR2, it was discovered by Helmi et al. (2018) that Gaia-Enceladus dominates the inner halo. Two distinct sequences are present in the color-magnitude diagram (CMD) of halo stars Gaia Collaboration et al. (2018). Halo stars are isolated by selecting stars with tangential velocities greater than 200 km/s. This implies two stellar populations are present, suggesting a so-called "dual" halo. The more metal-poor sequence consists largely of Gaia-Enceladus. The other sequence consists of stars that kinematically belong to the splash of the thick disk (fig. 2) Haywood et al. (2018). The accreted system is the result of a very massive object merging with the Milky Way around 10 Gyr ago.

1.5 Stellar streams in the Milky Way halo

One of the first substructures in the inner halo was discovered by Helmi et al. (1999), and subsequently got the name Helmi streams. The streams were identified through their kinematics. Due to their high z-velocities, they could be isolated from the other stars. The streams are thought to originate from a system with a mass of $M_* \sim 10^8 M_\odot$ (Koppelman et al., 2019).
Currently, it remains difficult to determine the origin of many substructures in the vicinity of the Sun. It is likely that better models and more detailed chemical analysis are needed on a large number of stars to accurately do this (Helmi, 2020). Recent works like Lövdal et al. (2022) and Ruiz-Lara et al. (2022) use data from the 6D sample of Gaia EDR3 in combination with spectroscopic data from amongst others, LAMOST and APOGEE to link clusters to larger substructures like Gaia-Enceladus and the Helmi streams.

A catalyst to the field of galactic archaeology was the discovery of the Sagittarius dwarf galaxy by Ibata et al. (1994), which is currently merging with the Milky Way. The Sagittarius streams are two branches consisting of stars that are tidally stripped from the core of the dwarf galaxy.

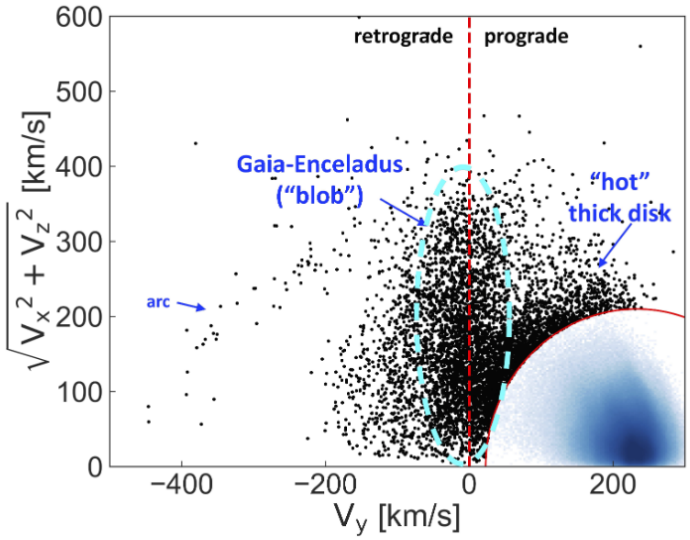


Figure 2: Toomre diagram, showing the velocity distribution of stars within 1 kpc of the Sun with data from the 6D sample of Gaia DR2. The blue density area in the bottom right are stars that belong to the disk. The others belong to the halo, which were selected as having tangential velocities greater than 210 km/s (Koppelman et al., 2018). Image from: Helmi (2020)

These streams make up one of the largest complex structures in the Milky Way halo, stretching from one side of the sky to the other (de Boer et al., 2015). Using photometric wide-field surveys like the Sloan Digital Sky Survey (SDSS), many substructures have been uncovered and identified in the outer halo of the Milky Way through their overdensities in the sky. It has allowed the structure of the Sagittarius streams to be defined (Ivezić et al., 2000)(Yanny et al., 2000). Another large substructure that has been discovered with data from the SDSS is the Virgo overdensity. This overdensity lies at a distance of $\sim$ 10 kpc, 4 times as close as the Sagittarius streams, however, both can be found in the same region on the sky (Jurić et al., 2008). Both substructures are thought to originate from dwarf galaxies merging with the Milky Way.

Many smaller thin streams have also been discovered in the outer halo, with digital sky surveys like SDSS, 2MASS, and WISE (Grillmair & Carlin, 2016) and the Dark Energy Survey (DES) (Shipp et al., 2018). These streams include GD-1, Orphan, Jhelum, Indus, and many others. Smaller streams like this, likely originate from less massive systems like globular clusters and small dwarf galaxies. The so-called "Field of Streams" from the SDSS data has allowed many streams to be discovered. The Sagittarius streams dominate a large part of the field. Figure 3 shows the "Field of Streams" around Orphan, named for its unknown progenitor (Belokurov et al., 2007). The color represents the distance and the intensity the density of stars.
New missions like Gaia have provided data that allows for the detection of substructures within stellar streams. E.g. a gap and a streak of off-stream stars in the stream GD-1 provide evidence of a dark satellite in the halo (Bonaca et al., 2019).

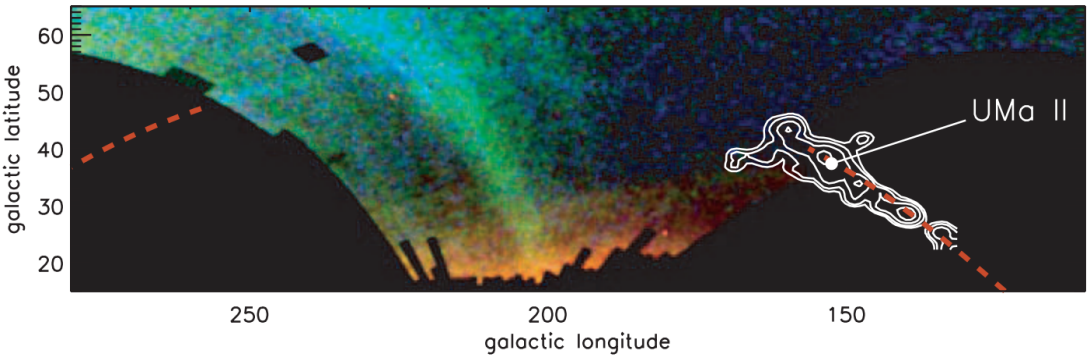


Figure 3: The "Field of Stream" from SDSS around the Orphan stream. The dashed red line is a fit of the orbit with a radius of $\sim$ 25 kpc around the galactic center. The bright and broad stream crossing Orphan is the Sagittarius Stream. The white contours correspond to HI column densities of the dwarf galaxies UMa II (Belokurov et al., 2007). However, the kinematics of the galaxy reveal that it is not the progenitor of Orphan Newberg et al. (2010)

.

1.6 Structure of the thesis

In this thesis, I will explore the possibilities of detecting stellar streams in the Milky Way halo using data from the Gaia early Data Release 3 (EDR3) (Gaia Collaboration et al., 2021). Specifically, I will use the reduced proper motion (RPM) selected halo sample created by Viswanathan et al. (2022). This sample provides photometric distances for $\sim$48 million stars, which are more accurate and precise than the parallax distances from Gaia and probes stars to higher distances. More details about Gaia and the Gaia EDR3 are provided in section 2.1. Information about the RPM halo sample and how it was created can be found in section 2.2.

The analysis consists of three main parts. In the first part, the distance to a selection of globular clusters will be determined. Section 2.4 will discuss the process of how stars belonging to a particular cluster are selected. Additionally, the photometric distance will be compared to distances computed using the metallicity and the effectiveness of using a color cut to remove stars with an incorrectly assigned absolute magnitude will be discussed. The distances to the GCs using the different methods are presented in section 3.1 and discussed in section 4.1.

The second part will explore the possibilities of creating an algorithm that can automatically find and isolate stellar streams by systematically probing different parts of the RPM selected halo sample. In section 2.5 the first steps are made in systematically going through the data. The structures found in this process are presented in section 3.2. Finally, section 4.3 will discuss the next possible steps in creating an algorithm to find stellar streams without being dependent on the 6D data sample of Gaia.

The final part of this thesis is focused on creating a selection of stars that belong to a stellar stream by using the overdensities in proper motion and the stream track, this process is described in more detail in section 2.6. The stellar stream chosen as a proof of concept for this selection is Phlegethon. Phlegethon is a stellar stream located nearby at an estimated distance of $\sim$ 3.8 kpc and covers a large range of proper motions (discovered by Ibata et al. (2018)). The final selection of the stellar stream and properties that can be inferred from it are presented in section 3.3. Section 4.2 will discuss the limitations and shortcomings of the results and the methods used.

2 Datasets and Methods

2.1 Gaia

Gaia was launched in December 2013, with its first data release in September 2016. The goal of Gaia is to gain knowledge about the composition, formation, and evolution of the Milky Way. Gaia has also revolutionized the detection of stellar streams and other substructures by providing accurate distances, proper motions, magnitudes, colors, and much more for hundreds of millions of stars. This has led to the discovery of new streams but also allows for verifying and defining known structures better. Future data releases will provide more complete radial-velocity catalogues and improve the precision of proper motions by a factor of 4.5 (ESA, 2020).

Malhan & Ibata (2018) have developed an algorithm by the name STREAMFINDER, that can be used to find thin stellar streams systematically and automated. It is able to find streams down to $> 10^\circ$ long, containing only ~ 15 members. The algorithm uses the proper motion and photometry of stars, and samples randomly over their radial velocities. By integrating orbits in a Galactic potential, stars that correspond to a thin stream-like structure in a certain orbit can be detected, allowing for the identification of a stellar stream. Gaia-1 and Gaia-2, are two of the streams that have been discovered using STREAMFINDER with data from the 6D sample of Gaia DR2 (Malhan et al., 2018). Many more streams have been added with the Gaia EDR3. These include Gaia-6 to Gaia-12 and twelve candidate streams (C-9 to C20) Ibata et al. (2021). The Gaia early Data Release 3 (EDR3) (Gaia Collaboration et al., 2021) contains 1.811 billion sources. 1.460 billion of these sources have full astrometric solutions.
A lot of research focused on stellar streams in the halo use the Gaia 6D sample, which contains the three-dimensional position and velocity. This sample is used to compute the integrals of motion. However, the 6D sample only contains 7.209 million sources. Most sources are missing the line-of-sight velocities, and these make up the Gaia 5D sample, i.e. the full Gaia EDR3 sample.

One method of computing the distance to stars is by taking the inverse of their parallax. However, around 90% of the stars in Gaia EDR3 have poor parallaxes due to their large error. Due to the absence of radial velocities and imprecise distances in the 5D sample, it is difficult to study the dynamics of the stellar halo. However, the Gaia EDR3 does have reliable astrometric and photometric measurements. This information can be used to calculate more accurate distances using the reduced proper motion while still probing Gaia's faintest limit.

2.2 Reduced Proper Motion selected halo sample

An alternative to using parallax distances is to calculate the photometric distances of stars based on their apparent magnitude.

$$M_G = m_G - 5\log_{10}\left(\frac{d}{\text{kpc}}\right) - 10 \tag{1}$$

Where M_G and m_G are the absolute and apparent magnitude in the Gaia G-band respectively, d is the distance. M_G is known for "standard candle" stars like cepheids and RR Lyrae Bhardwaj (2018). For main-sequence (MS) stars there is an almost linear relation between the apparent magnitude and colour, which can be used to determine M_G for MS stars. This process of determining the photometric distances is described by Koppelman & Helmi (2021) and applied to the stars in Gaia EDR3 by Viswanathan et al. (2022).

The main sequence stars are selected using the reduced proper motion (RPM) parameter, based on the photometry and proper motions. The reduced proper motion is defined by:

$$H_G = m_G - 5\log\mu - 10 \tag{2}$$

Where μ is the Gaia proper motion. Based on the proper motions μ_j and the photometric distances, the tangential velocities can be determined using

$$v_j = 4.740\,57\,\mathrm{km/s}\left(\frac{\mu_j}{\mathrm{mas/yr}}\right)\left(\frac{d}{\mathrm{kpc}}\right) \tag{3}$$

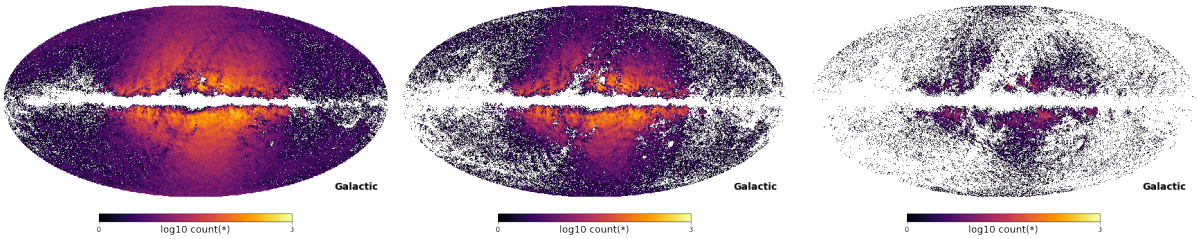
Where $j = (l, b)$. The factor 4.74057 results from the unit conversion to km/s.
Using eqs. (1) and (3), the following expression for the reduced proper motion can be obtained from eq. (2):

$$H_G = M_G - 5\log\frac{v_{tan}}{4.74057} \tag{4}$$

Where v_{tan} is the tangential velocity. All these magnitudes are extinction corrected (see Viswanathan et al. (2022)). Using the dependence of H_G on the absolute magnitude, MS stars can be selected using the H_G-colours diagram. This RPM diagram is equivalent to the HR diagram of a stellar population with an offset due to v_{tan}. Stars in the halo orbiting the galactic center typically have high tangential velocities compared to the disk. Therefore, halo stars are separated from stars from the disk by selecting high v_{tan} stars, which show up as a separate sequence in the RPM diagram. This method can reliably select halo stars, it is however not complete. Stars that move along the line of sight will have a small v_{tan} and are therefore excluded. After selecting stars in the MS region of the RPM diagram, with high v_{tan}, an estimate can be made for M_G, based on the star in the position in the RPM diagram. Photometric distances for these MS stars can be determined using their linear relation between the absolute magnitude and color. Stars outside the MS, like the giants and stars at the MS turn-off, don't have a reliable enough dependence on the absolute magnitude and color (Koppelman & Helmi, 2021).

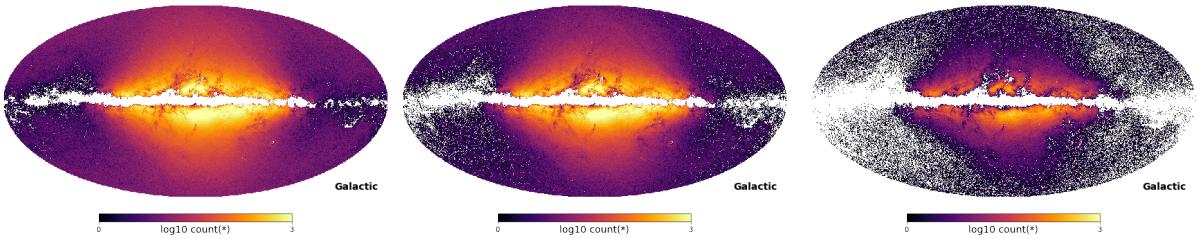
As a quality cut, only stars with an uncertainty in the reduced proper motion parameter $\frac{H_G}{\delta H_G} > 1.75$ are selected. This results in a final sample of around 48 million stars. Additionally, a colour cut can be made to exclude turn-off and redder stars: $0.45 > G - G_{RP} > 0.715$, leaving around 25 million stars. This sample contains around 3.5 times as many stars, as the equivalent sample produced from Gaia DR2 by Koppelman & Helmi (2021).

Additionally, compared to RPM selected halo sample of Gaia Data Release 2 (DR2) (Gaia Collaboration et al., 2018), Gaia EDR3 populates the halo much more, especially at larger distances. This can been clearly seen in figs. 4 and 5, which show the number density of Gaia DR2 and EDR3 respectively, for different distances for the RPM selected halo stars (see section 2.2). The average photometric distance in the RPM selected halo sample is 4.3 kpc for Gaia DR2, compared to 8kpc for Gaia EDR3 (Viswanathan et al., 2022). The spatial distribution in cylindrical heliocentric space (fig. 6), shows the sample populates space up to 15 kpc. A similar figure in Koppelman & Helmi (2021) shows the RPM selected halo sample of Gaia DR2 populates space well up to ~ 8 kpc. Additionally, Gaia EDR3 improves the proper motion accuracy by a factor of 2, and parallaxes by a factor of 1.5 compared to Gaia DR2.



(a) $d > 0$ kpc (b) $d > 5$ kpc (c) $d > 10$ kpc

Figure 4: Galactic map of the RPM selected halo stars of Gaia DR2 at (a) all distances, (b) photometric distances greater than 5 kpc and (c) photometric distances greater than 10 kpc.



(a) $d > 0$ kpc (b) $d > 5$ kpc (c) $d > 10$ kpc

Figure 5: Galactic map of the RPM selected halo stars of Gaia EDR3 at (a) all distances, (b) photometric distances greater than 5 kpc and (c) photometric distances greater than 10 kpc.

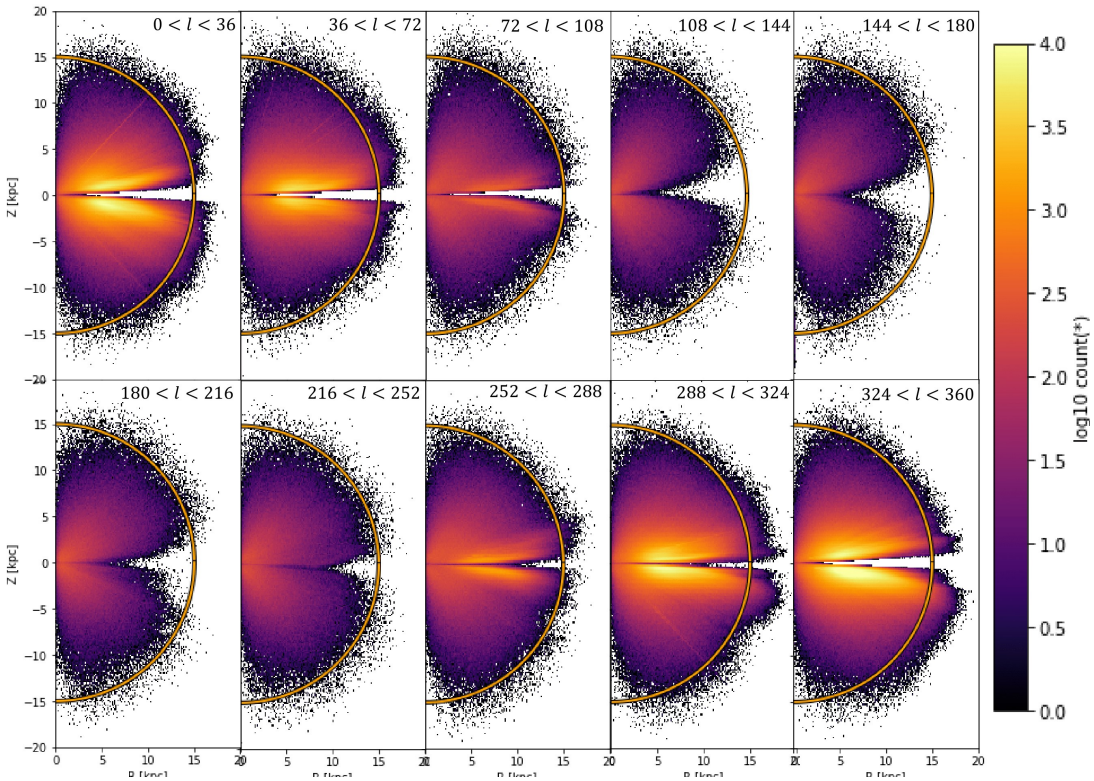


Figure 6: Spatial distribution of the RPM selected halo sample in cylindrical heliocentric coordinates for different galactic longitude sections. A half circle at 15 kpc around the origin is drawn for comparison

2.3 Substructures in velocity space

One simple method of looking for substructures is through velocity vectors. Structures can be detected if their velocity is distinctly different enough from the background halo.
To obtain the right velocities in the galactic coordinate space, the velocities determined using eq. (3), have to be corrected for the solar motion. The correction factors are defined by:

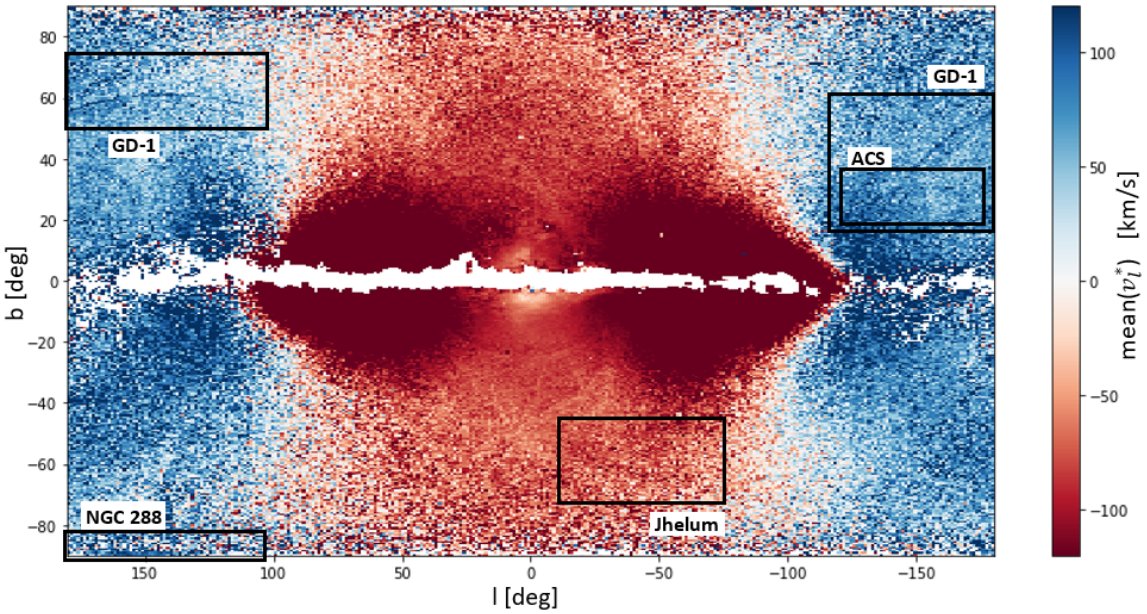
$$v_{l,\odot} = -U_\odot \sin l + (V_\odot + V_{LSR}) \cos l \tag{5}$$

$$v_{b,\odot} = W_\odot \cos b - \sin b(U_\odot + (V_\odot + V_{LSR}) \sin l) \tag{6}$$

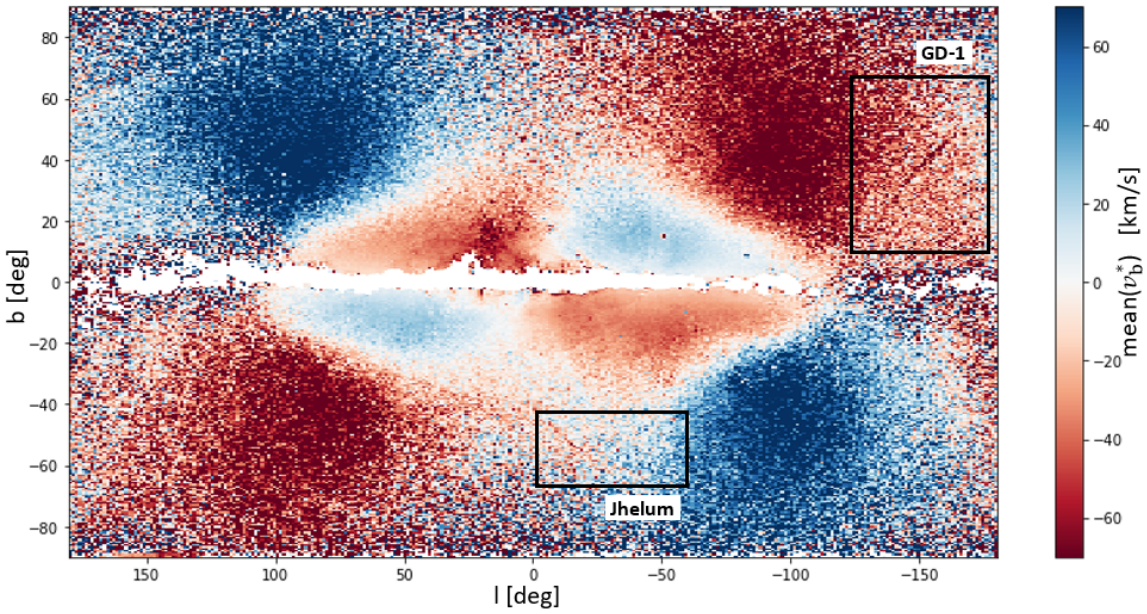
where the solar motion is defined as $(U_\odot, V_\odot, W_\odot) = (11.1, 12.24, 7.25)$ km/s (Schönrich et al., 2010) and the local standard of rest motion $V_{LSR} = 232.8$ km/s (McMillan, 2017). Finally, the correction is added to the velocity determined with eq. (3).

$$v_i^* = v_i + v_{i,\odot} \tag{7}$$

Figures 7a and 7b show that several substructures and overdensities can be observed in binned velocity space for distances greater than 7 kpc which is around where the mean of the sample lies.



(a) Binned longitudinal velocities corrected for solar motion



(b) Binned latitudinal velocities corrected for sollar motion

Figure 7: Galactic maps of distances greater than 7 kpc, with (a) binned longitudinal velocities and (b) latitudinal velocities, that are corrected for solar motion. Several streams and substructures can be observed, a number of them are annotated.

Note: All data analysis was performed using Python. In this section, it will frequently be mentioned a fit to a Gaussian or polynomial function was performed. This was done using `scipy.optimize.curve_fit`*, which performs a non-linear least-squares fit (Virtanen et al., 2020). The python module* `vaex` *(Breddels & Veljanoski, 2018), allows to efficiently use and explore large tabular datasets. It was also used to visualize and do computations on the data. The main components of Python code used for the computations of each part of the analysis can be found in appendix B.*

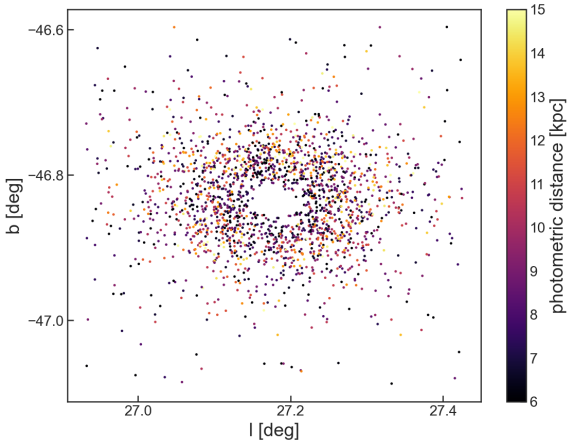
2.4 Distance to Globular Clusters

When looking through the EDR3 density plots and different proper motions, described in section 2.5, multiple globular clusters (GCs) can easily be identified. These clusters have known distances and thus can be used to verify the photometric distances of the RPM selected halo sample. The stars belonging to these GCs can be isolated through their overdensity due to proper motion selections. GCs are identified through their galactic coordinates and validation is done by looking at a GC database by (Massari et al., 2019).

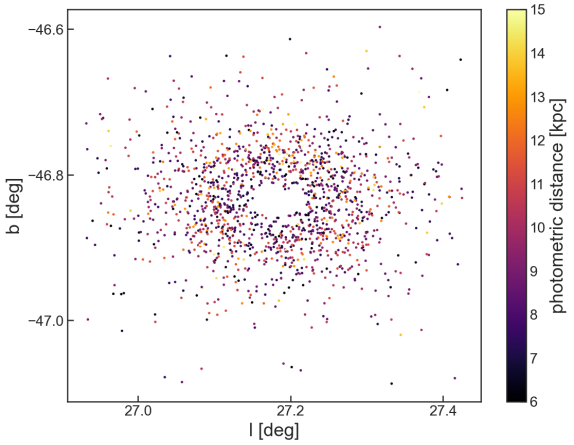
Using the galactic coordinates corresponding to the center of the GCs, a square area of 0.5 x 0.5 degrees around the center of each GC is selected (fig. 8a). After determining the peak of the binned proper motion space in μ_l, all stars within 1.3σ around the peak are selected (fig. 8b). This range allows the majority of stars in the peak to be selected, while also accounting for the errors in proper motion. However, the proper motion selection to isolate the stars corresponding to the GC is not always successful. Therefore a Gaussian is fit to the number of stars along the galactic longitude l. All stars within the full width at half maximum (FWHM) around the center are selected (fig. 8c). Finally, the distance and its dispersion are determined by taking the mean and standard deviation of the photometric distances of the selected stars respectively.

The colour cut described in section 2.2 ($0.45 > G - G_{RP} > 0.715$) can be applied to remove very blue stars (fig. 8d). Stars in this range have colors close to the MS turn-off. Consequently, their absolute magnitude would be incorrectly assigned, leading to an underestimated distance.

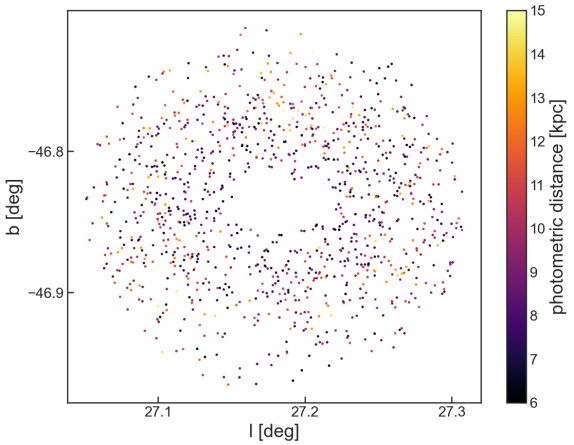
Additionally, the photometric distances are compared to the distances computed using the metallicity of the GCs compiled by Massari et al. (2019). The absolute magnitude M_G is determined based on interpolation of the relationship between M_G and $G - G_{RP}$ for a given metallicity, given by Viswanathan et al. (2022). The distance can then be calculated using the distance modulus (eq. (1)).



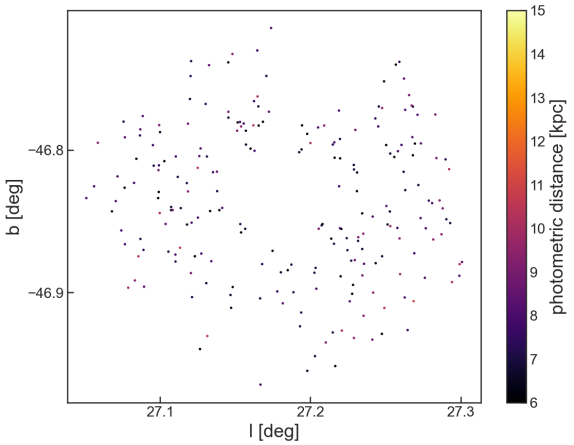
(a) All stars with 0.5° around the center of the GC



(b) Stars around the peak in proper motion in fig. 8a



(c) Stars within the FWHM of the density distribution



(d) After colour cut: $0.45 > G - G_{RP} > 0.715$

Figure 8: An example of the selection process of GCs. The GC shown here is NGC7099

2.5 Systematic search for substructures

To effectively look for substructures in the Milky Way halo, an algorithm must be developed that automatically and systematically looks for structures and potentially makes a selection of these stars. Unlike STREAMFINDER, this method would use the 5D sample to not be limited by the small number of sources that have line-of-sight velocities.

A first step in the direction of such an algorithm was made by slicing in the galactic coordinates, proper motion space, and distance. By systematically probing the sky like this, overdensities will become apparent. To test the method a set of scatter plots were created for 60° by 60° areas in galactic coordinates for proper motion bins in μ_l and μ_b with a 5 mas/yr width for distances greater than 8 kpc. Low latitude sources ($|b| < 30°$) were excluded due to their high contamination by stars from the thick disc. Only selections with over 2000 sources were plotted. The selection of potential stellar streams and other substructures was done by eye. Six of the streams that were found and identified using Ibata et al. (2021), can be seen in section 3.2.1.

2.6 Star selection of the stellar stream Phlegethon

The stellar stream chosen to be examined in more detail is Phlegethon. Phlegethon was discovered by Ibata et al. (2018) using the STREAMFINDER algorithm on data from Gaia DR2. The stream is extremely retrograde and lies at an estimated distance of ≈ 3.8 kpc. The stream is thought to be a remnant of a globular cluster. Because Phlegethon is located in the solar neighborhood and is difficult to isolate in binned velocity space, it was unexplored in Viswanathan et al. (2022). The stream is also interesting due to its location close to the galactic plane and low surface brightness. Therefore showing the used method of star selection is reliable for Phlegethon, will be a good proof of concept that the method is likely to be effective on stellar streams in the outer halo too and/or with higher surface brightness. Additionally, STREAMFINDER is only able to detect thin streams. It is possible Phlegethon is broader and has more gaps, wiggle, and density variations than that have previously been reported. With the RPM sample, we can look more into this.

Based on a rough estimation of the area, proper motion space, and distance of Phlegethon, a density plot with Gaussian smoothing (fig. 9) was created. The distance was limited to be greater than 2.5 kpc and smaller than 4.5 kpc. The galactic coordinates were limited an area where $-20° < l < 70°$ and $-50° < b < -35°$. The proper motion corrected for the solar motion was bounded by $-40 < \mu_l^* < -20$ mas/yr and $-30 < \mu_b^* < 5$ mas/yr. These boundary values were partly determined by experimenting with which combination of selections showed the clearest overdensity and partly on information provided by Ibata et al. (2018). Stars in the higher density regions of the density map were isolated to make an initial selection of stars.

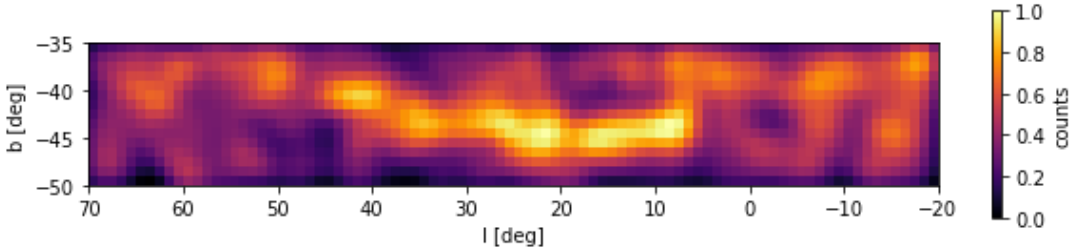


Figure 9: Normalized star density of area around Phlegethon with proper motions limited to $-40 < \mu_l^* < -20$ mas/yr and $-30 < \mu_b^* < 5$ mas/yr and distances to $2.5 < d < 4.5$ kpc.

This selection was used to fit a 2nd-degree polynomial, representing the orbit. All stars $\pm 8°$ in galactic latitude b were selected (eq. (8)). The previously set boundaries were expanded to: $2.5 < d < 5$ kpc, $-40 < \mu_l^* < -9$ mas/yr and $-35 < \mu_b^* < 17$ mas/yr. No limits were set on the galactic coordinates. The relatively large width of $8°$ was chosen because of large inaccuracy in this initial fit to the orbit due to the still large amount of contamination in the sample.

$$b < (0.00315\, l^2 - 0.134\, l - 41.9) \pm 8° \tag{8}$$

Secondly, Phlegethon was isolated using proper motion vectors. Proper motion vectors were used as opposed to velocity vectors since the latter introduces additional uncertainties. The photometric distances in the RPM selected halo sample have poor accuracy for MS turn-off stars and fainter redder stars.

First, a 3rd degree polynomial was fitted to the overdensity in l vs. μ_b (fig. 10a).The fit was performed on selected points near the overdensity, opposed to all sources in the selected sample. All stars within ten standard deviations σ from the mean in the error in μ_b were selected:

$$\mu_b < (3.8 \times 10^{-5}\, l^3 - 4.3 \times 10^{-5}\, l^2 - 0.51\, l + 5.0) \pm \left(\bar{\delta(\mu_b)} + 10\sigma\delta(\mu_b)\right) \quad \text{[mas/yr]} \tag{9}$$

Based on this selection, an improved fit of the stream-track and corresponding selection was created by fitting a 2nd degree polynomial to selected points in the spatial overdensity. Subsequently, a 2nd degree polynomial is fit to the overdensity in l vs. μ_l (fig. 10b).

$$\mu_l < (9.3 \times 10^{-3}\, l^2 - 0.32\, l - 35.2) \pm \left(\bar{\delta(\mu_l)} + 6\sigma\delta(\mu_l)\right) \quad \text{[mas/yr]} \tag{10}$$

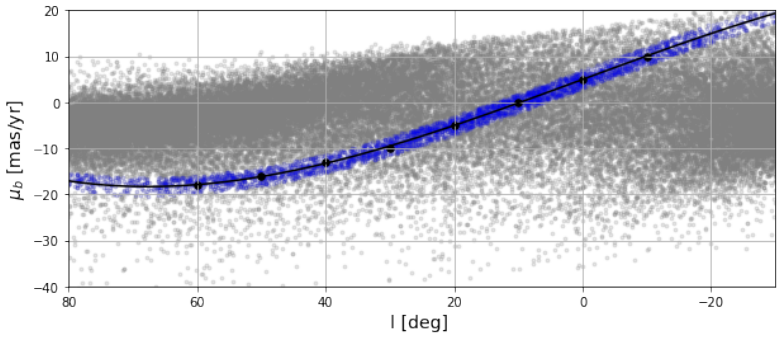
This was followed by another improved fit and selection of the stream track. However, now enough contamination had been removed to let the points on the overdensity be automatically selected. The current selection sample was put into bins with an $\Delta l = 10^\circ$ width, from which the median of b was taken. The improved fit of the stream track is followed by the selection of the overdensity in l vs. μ_{RA} (fig. 10c). For this fit, the data has again been put into bins of $\Delta l = 10^\circ$. By taking the peak of the histogram in μ_{RA} for every bin, points were created to fit the polynomial.

These methods were repeated for another improved fit of the stream track and for the overdensity in μ_{DEC} vs. l (fig. 10d), respectively.The final fit was performed on all data points in the selection. For every new fit to the track, the width of the selection around the track was decreased from 8° for the first to 4° for the final selection. The final fit resulted in the following polynomial stream track:

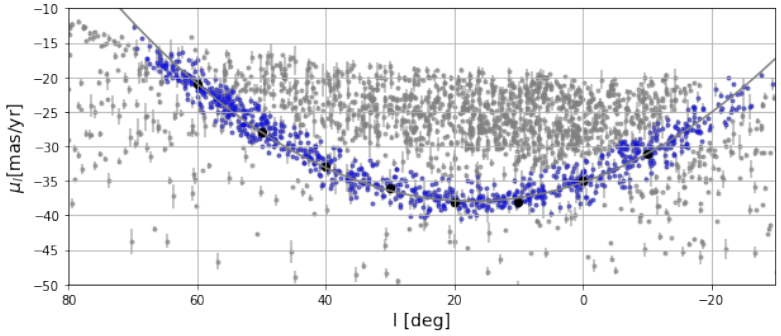
$$b(l) = (5.8 \pm 0.5) \times 10^{-3}\, l^2 - (0.22 \pm 0.02)\, l - (42.4 \pm 0.2)$$

Where l and b are given in degrees.

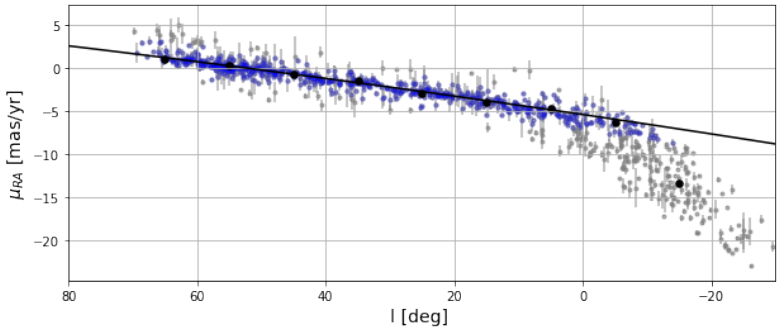
In the end, the selection consisted of 575 potential members of Phlegethon. A corresponding density map of Phlegethon is created to better analyze its substructure.



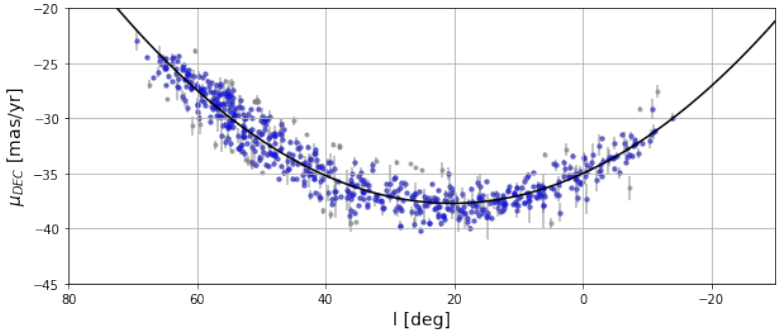
(a) $\mu_b < (3.8 \times 10^{-5}\ l^3 - 4.3 \times 10^{-5}\ l^2 - 0.51\ l + 5.0) \pm \left(\bar{\delta(\mu_b)} + 10\sigma\delta(\mu_b)\right)$ [mas/yr]



(b) $\mu_l < (9.3 \times 10^{-3}\ l^2 - 0.32\ l - 35.2) \pm \left(\bar{\delta(\mu_l)} + 7\sigma\delta(\mu_l)\right)$ [mas/yr]



(c) $\mu_{RA} < (-1.3 \times 10^{-4}\ l^2 - 0.11\ l - 5.4) \pm \left(\bar{\delta(\mu_{RA})} + 3\sigma\delta(\mu_{RA})\right)$ [mas/yr]



(d) $\mu_{DEC} < (6.5 \times 10^{-3}\ l^2 - 0.27\ l - 35.0) \pm \left(\bar{\delta(\mu_{DEC})} + 6\sigma\delta(\mu_{DEC})\right)$ [mas/yr]

Figure 10: The selection of members of the stellar stream Phlegethon through 4 proper motions vectors: (a) μ_b, (b) μ_l, (c) μ_{RA}, (d) μ_{DEC}. A polynomial is fit to selected points in the overdensity present in each plot. Stars around the polynomial fit are selected and shown in blue.

3 Results

3.1 Distance to globular clusters

In total, eight globular clusters were chosen for the distance analysis. Table 1 shows the mean photometric distance and mean metallicity-dependent distance for the data selections without the color cut, and the distances from Massari et al. (2019). The GCs catalogue did not contain a metallicity for NGC6402, hence no metallicity-dependent distance could be determined for this GC.

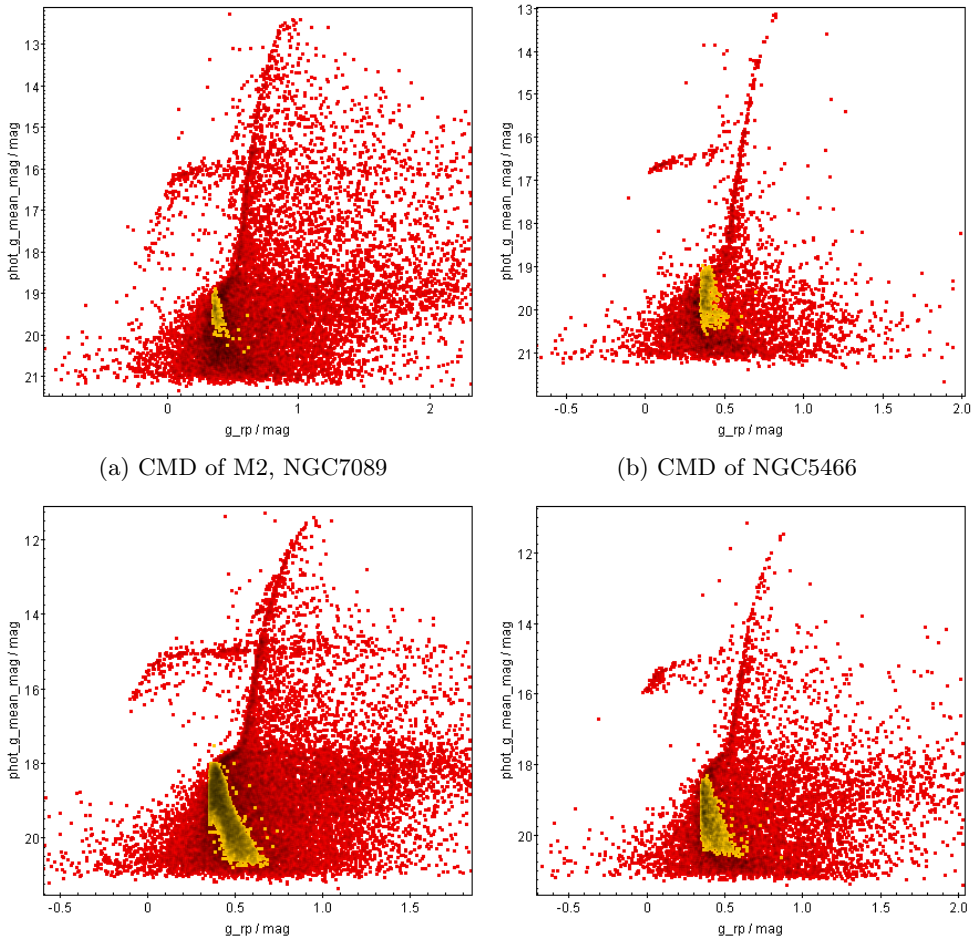
Object	**Phot. Dist. [kpc]**	**Met. dep. Dist. [kpc]**	**Lit. value [kpc]**
M2, NGC7089	10.1 ± 1.5	8.2 ± 1.9	11.50
M5, NGC5904	7.3 ± 1.4	7.9 ± 1.4	7.50
M13, NGC6205	11.0 ± 1.4	10.4 ± 1.3	7.10
M14, NGC6402	8.4 ± 1.7	N/A	9.30
M30, NGC7099	9.3 ± 2.2	7.7 ± 1.7	8.10
M80, NGC6093	8.8 ± 1.6	7.5 ± 1.2	10.0
M92, NGC6341	10.5 ± 2.0	8.7 ± 1.4	8.30
NGC5466	11.1 ± 2.1	8.8 ± 1.8	16.0

Table 1: The computed mean photometric and metallicity dependent distances for eight globular clusters, without the color cut. The literature values in the last column are obtained from Massari et al. (2019). GCs with reliable (metallicity dependent) distances are shown in bold.

To determine which combination of sample and method provides the most accurate distances, two plots were created showing deviation from literature for both the photometric and metallicity determined distance method for the data sample without the color cut (fig. 12a) and with the color cut: $0.45 > G - G_{RP} > 0.715$ (fig. 12b).

When the color cut is used, the number of stars in the selection of most GCs greatly decreases. This is likely because many stars in these GCs belong to the MS turn-off, consequently, their absolute magnitude would be incorrectly assigned, leading to an underestimated distance. We can check this by superimposing the GC selection of the RPM selected halo sample onto a rough selection of the GC in the full Gaia EDR3 sample. This shows that amongst others, the selections of NGC7089 and NGC5466 are mostly part of the MS-turnoff (figs. 11a and 11b), whereas the selections of NGC5904 and NGC7099 do extent into the MS-turnoff, they still cover the MS (figs. 11c and 11d). All sky distributions and CMDs of the GCs can be found in appendix A.

In total, for four out of the seven GCs, the stars in the selection belonged to the MS turn-off. Only three GCs have more than 100 stars left after the color cut: NGC6341, NGC7099, and NGC5904. Subsequently, these are the only GCs for which the deviation from literature is around zero. From this, we can conclude that the color cut allows us to distinguish accurate distances from incorrect ones. After the color cut, there is little difference in accuracy between the photometric and metallicity determined distances. However, without the color cut, the metallicity-dependent distances for GCs with correctly assigned magnitudes are more accurate.



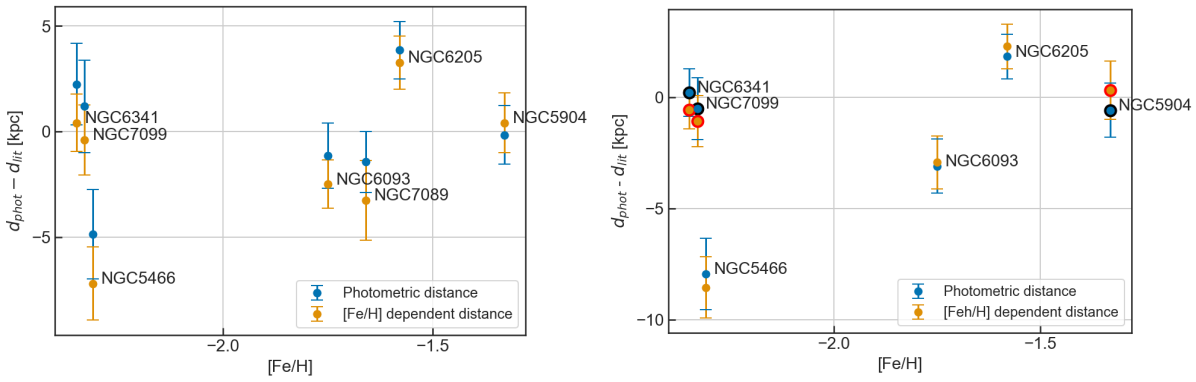
(a) CMD of M2, NGC7089

(b) CMD of NGC5466

(c) CMD of M5, NGC5904

(d) CMD of M30, 7099

Figure 11: The GC selections of the RPM selected halo sample without the color-cut (yellow) are superimposed onto a rough selection of the GC for the full Gaia EDR3 sample (red) (all stars with 0.15° of the GC's center). The selection of NGC7089 and NGC5466 consist of mostly MS-turnoff stars. The selections of NGC5904 and NGC7099 do cover the MS.



(a) No colour cut

(b) With colour cut: $0.45 > G - G_{RP} > 0.715$

Figure 12: Blue: difference between the photometric distance and literature distance (Massari et al., 2019). Orange: difference between metallicity-dependent distance and literature distance. In (b), GCs for which >100 stars were left after the color cut are shown with a larger marker.

3.2 Systematic selection substructures

3.2.1 Identified stellar streams

In total six known stellar streams were located by eye using the scatter plots created by looping over different areas and proper motions. The streams that were found are: GD-1, Jhelum, Indus, Ophiuchus, Ylgr, and Orphan. Their corresponding plots from the systematic search are shown in figs. 13 to 19. The proper motion bins can be found in the figure caption, and the spatial bins are represented on the axes. For every plot the data is colour-coded for the solar motion corrected longitudinal proper motion μ_l^*.

GD-1

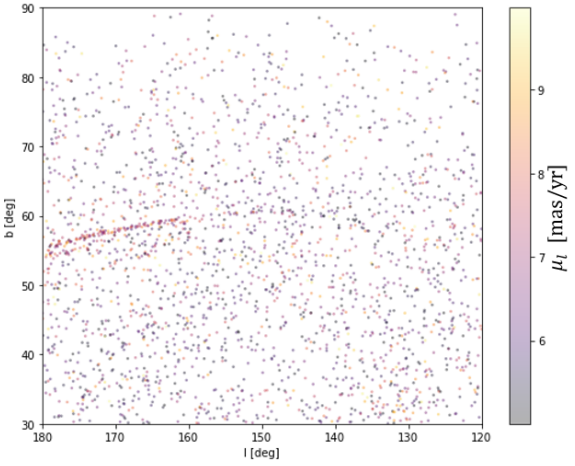


Figure 13: $5 < \mu_l^* < 10, -5 < \mu_b^* < 0$ [mas/yr] Scatter plot for the shown proper motion and spatial interval, and $d_{phot} > 8$ kpc. The data is colour-coded for the solar motion corrected longitudinal proper motion μ_l^*.

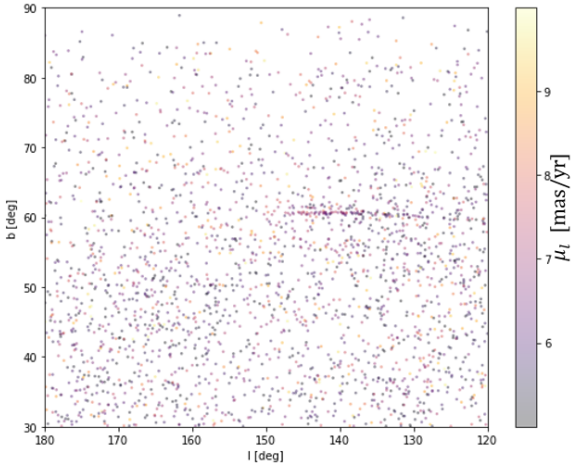


Figure 14: Same as fig. 13 but for $5 < \mu_l^* < 10, 0 < \mu_b^* < 5$ [mas/yr]

Jhelum and Indus

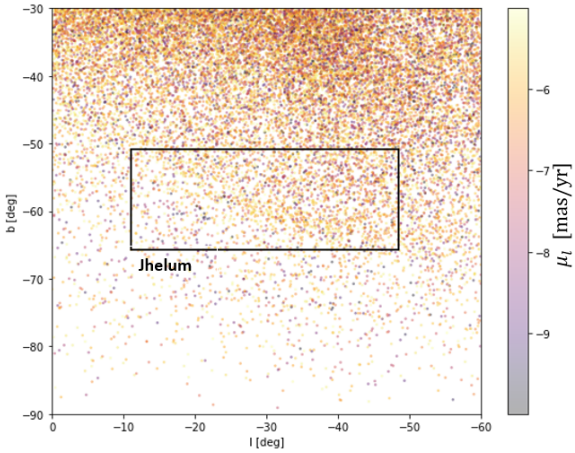


Figure 15: Same as fig. 13 but for $-10 < \mu_l^* < -5, -5 < \mu_b^* < 0$ [mas/yr]

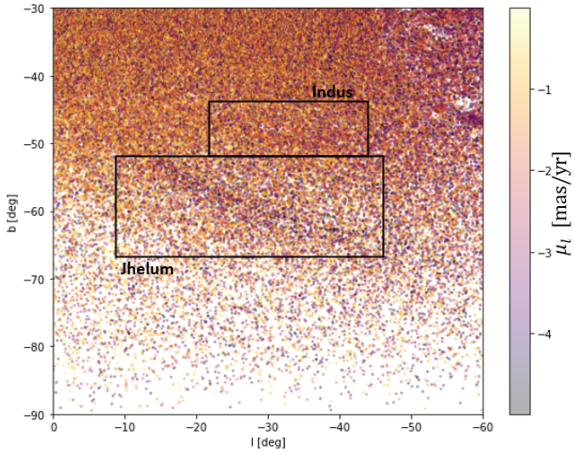


Figure 16: Same as fig. 13 but for $-5 < \mu_l^* < 0, -5 < \mu_b^* < 0$ [mas/yr]

Ophiuchus

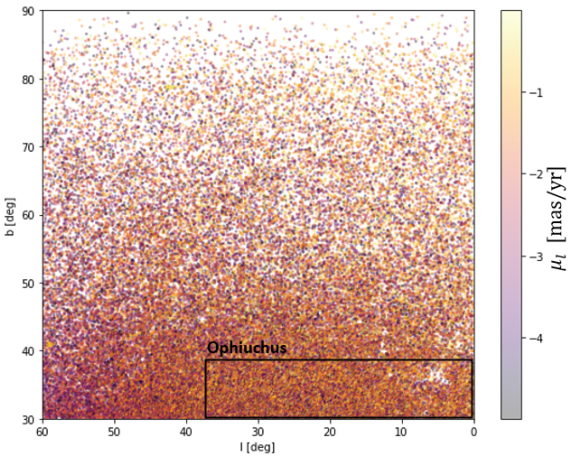


Figure 17: Same as fig. 13 but for $-5 < \mu_l* < 0, -5 < \mu_b^* < 0$ [mas/yr]

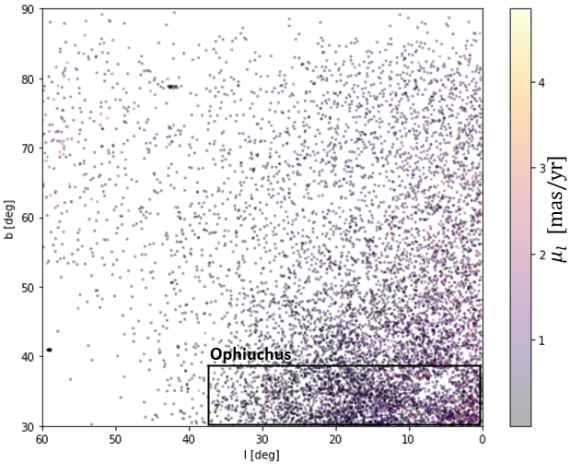


Figure 18: Same as fig. 13 but for $0 < \mu_l^* < 5, -5 < \mu_b^* < 0$ [mas/yr]

Ylgr and Oprhan

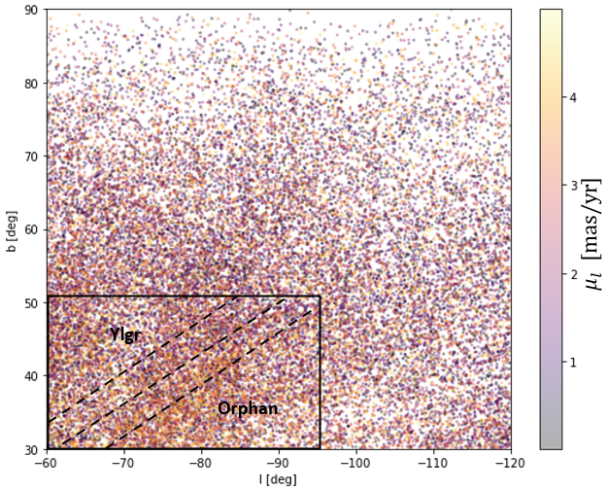


Figure 19: Same as fig. 13 but for $0 < \mu_l^* < 5, -5 < \mu_b^* < 0$ [mas/yr]

In the systematic search for stellar streams, only structures with $|b| > 30°$ and distances greater than 8 kpc were considered. However, to assess the possibility of detecting nearby streams closer to the galactic plane with a similar method, specific stellar streams were identified using a rough estimation of their proper motions based on Ibata et al. (2021). Two of the stellar streams that were identified are Hríd and Gjöll using $15 < \mu_l^* < 25$ mas/yr, $7 < \mu_b^* < 25$ mas/yr and distances between 3 and 6 kpc. Figure 20 shows that both streams are resolved using these boundary parameters. Similarly, the stellar stream Phlegethon could also be identified like this.

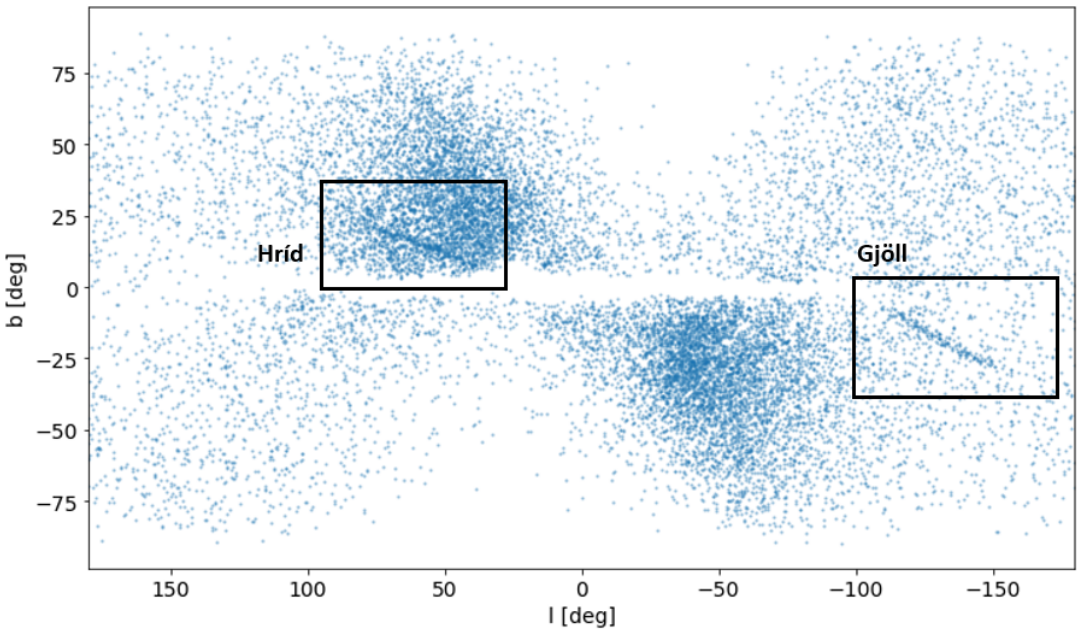


Figure 20: Galactic map showing overdensities who have identified as the stellar streams Hríd and Gjöll. Stars with proper motions $15 < \mu_l^* < 25$ mas/yr, $7 < \mu_b^* < 25$ mas/yr and distances between 3 and 6 kpc have been selected from the RPM selected halo sample for this plot.

3.2.2 Scanning pattern

Additionally, some of the substructures that could not be identified were found to be a result of the scanning pattern of Gaia. Using the `astrometric_n_obs_al` parameter, which shows the total number of observations for sources, the scanning pattern can be revealed. This is shown in gray scale in the figures. By superimposing the sources of a single plot created in the systematic search for substructures on the scanning pattern, one can observe whether there is a correlation between the two. A few examples of substructures due to the scanning pattern are shown in figs. 21 and 22

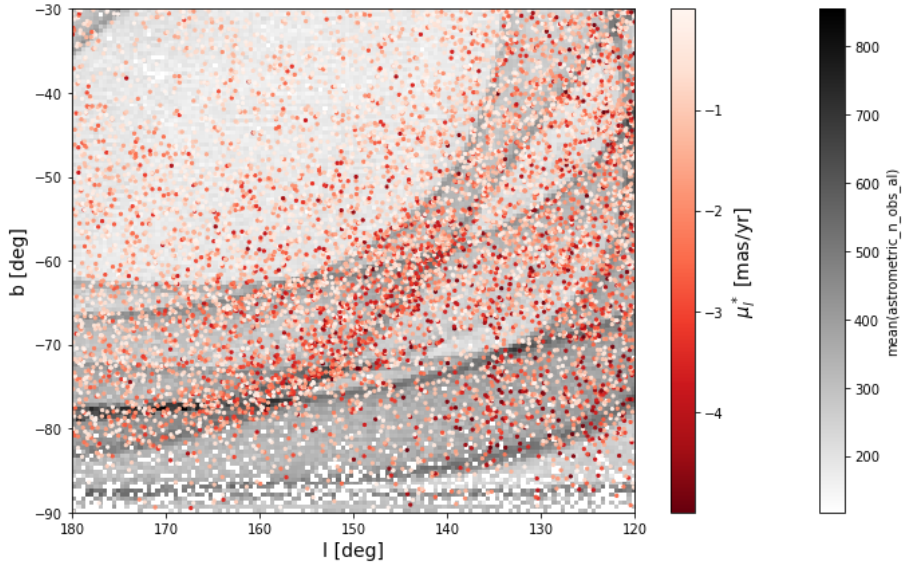


Figure 21: Galactic map of sources with proper motions $-5 < \mu_l^* < 0, -5 < \mu_b^* < 0$ [mas/yr] and distances $8 < d < 25$ kpc (red) superimposed on the scanning pattern (grey).

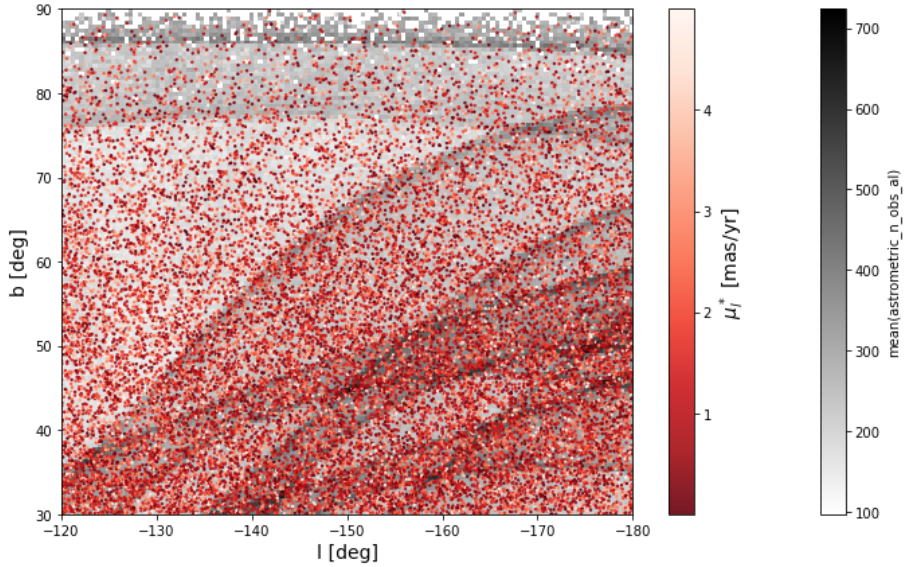
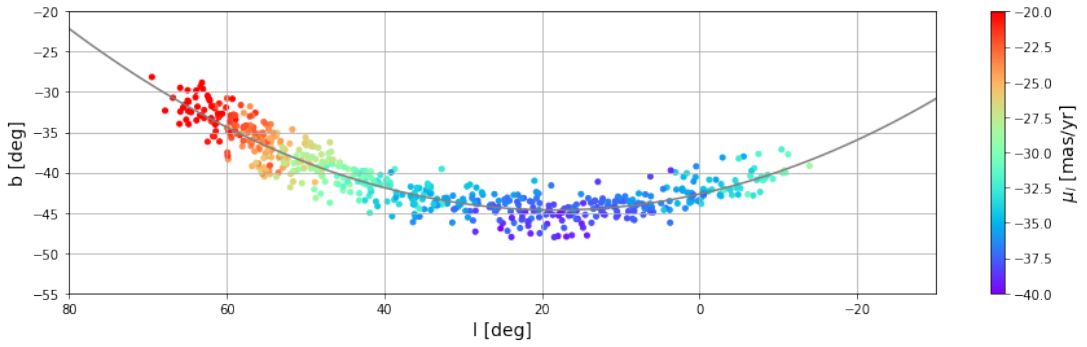


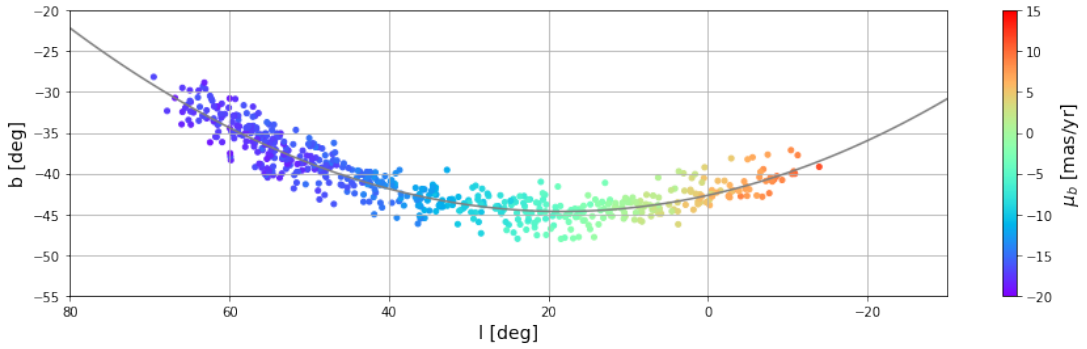
Figure 22: Galactic map of sources with proper motions $0 < \mu_l^* < 5, -5 < \mu_b^* < 0$ [mas/yr] and distances $8 < d < 25$ kpc (red) superimposed on the scanning pattern (grey).

3.3 Analysis of the Phlegethon selection

The sequence of selections on proper motion overdensities and the stream-track results in a final selection of 575 stars. Figure 23 shows the distribution of the stars in galactic coordinates, which show a clear stream-like structure. The colored dots reveal a gradient in both μ_l and μ_b. Additionally, fig. 24 shows the distribution in proper motion, resembling an curvature in the continuous structure. There is a large gradient in the proper motions, with $-40 < \mu_l < -15$ and $-20 < \mu_l < 15$. The colored dots show a continuous gradient in the galactic longitude. The continuity in the three parameters corresponds to what is expected for structures like stellar streams.



(a) Sky distribution with latitudinal proper motion.



(b) Sky distribution with longitudinal proper motion.

Figure 23: The distribution in galactic coordinates of the 575 selected stars of Phlegethon. The line is the best polynomial fit to the data, representing the track the stream follows. The data is color-coded for (a) μ_l and (b) μ_b.

The mean photometric distance of the selected Phlegethon sample is 3.6 ± 0.6 kpc. The large uncertainty is due to the large distribution of distances in the sample. Based on the distance of the stream and its angular size, its true size can be determined. Assuming a length of $\sim 80°$, its physical length is ~ 5 kpc, and a width of $\sim 5°$ corresponds to ~ 0.3 kpc.

The proper motion μ_l is negative throughout the whole structure, indicating that the stream moves in the negative l direction (towards the right in fig. 23a) and his highly retrograde. μ_l is lowest around $l = 18°$, where $\mu_l \approx -40$ mas/yr. μ_b is lowest at high l end of the stream at $l \approx 70°$, where $\mu_b \approx -20$ mas/yr and highest and the low l end ($l \approx 70°$) where $\mu_b \approx 10$ mas/yr. To make the substructure of Phlegethon more apparent, a density plot was created (fig. 25). This reveals a larger density on the high l end and a relatively lower density on the low l end

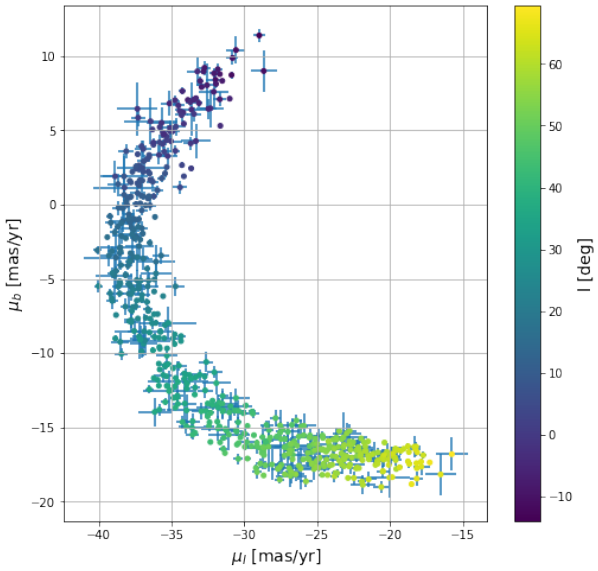


Figure 24: Proper motion distribution for the selected Phlegethon sample. The colour represent the galactic longitude, showing a continuous gradient through the structure. The uncertainties in the individual points are shown through errorbars.

of the stream. In combination with the movement direction and proper motions gradient, this indicates that the stream is growing longer as it is being pulled apart further by tidal forces.

A gap in the stream can be observed around $l = 40°$, which is also visible in fig. 24 at $(\mu_l, \mu_b) \approx (-35, -10)$ mas/yr. A smaller gap is present around $l = 3°$, corresponding to $(\mu_l, \mu_b) \approx (-36, 3)$ mas/yr. Additionally, possible parallel structures can be seen below the the stream around $l \sim 55°$ and above the stream around $l \sim -5°$. They are however very faint, due to not being fully resolved, as the selection was focused on Phlegethon.

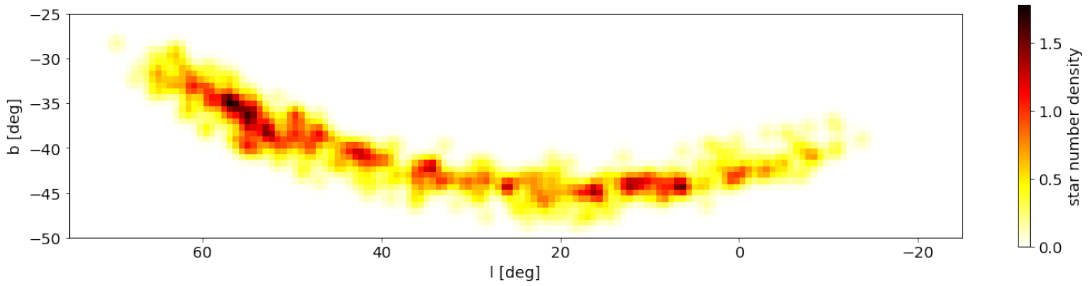


Figure 25: Galactic density map of the Phlegethon selection.

The stars in the selected Phlegethon sample can be cross-matched with Pristine (Martin et al., 2022), resulting in 38 cross matches. The distribution of the cross matches is displayed in fig. 26, showing all stars have a galactic longitude greater than 55°. This is a consequence of the footprint of Pristine. The metallicity distribution is shown through a histogram in fig. 27. A Gaussian was fit to this histogram, from which was determined that the mean spectroscopic metallicity of these stars is [Fe/H] $= -1.9 \pm 0.5$, which corresponds to the very metal-poor range. The metallicity range is relatively broad compared to other streams originating from GCs Martin

et al. (2022), ranging from [Fe/H] ~ -1.0 to [Fe/H] ~ -3.0. This could indicate that the stream is instead part of an accreted dwarf galaxy. However, some contamination and large metallicity uncertainties for faint stars could also have influenced the large dispersion.

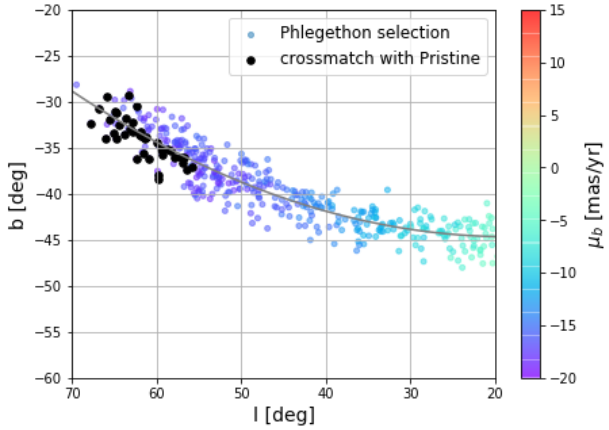


Figure 26: The distribution of the cross matches between the Phlegethon selection and Pristine in galactic coordinates. 38 cross matches were found in total

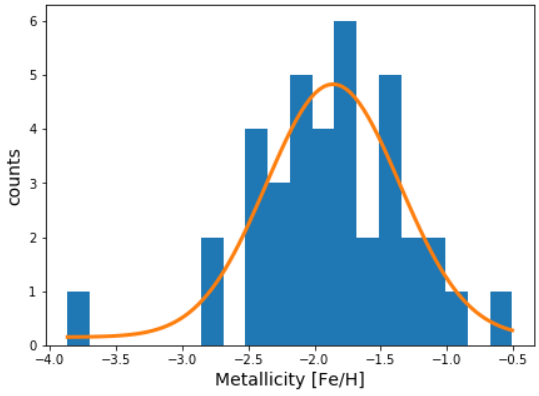


Figure 27: Histogram of the metallicities in the cross-match of the Phlegethon selection with Pristine. Based on the Gaussian fit to the histogram, the mean metallicity is [Fe/H] $= -1.9 \pm 0.5$.

4 Discussion

4.1 Distance to globular clusters

The uncertainties in the distance without the color cut ($0.45 > G - G_{RP} > 0.715$) are relatively big, with a mean error of 1.7 kpc for the photometric distances and 1.5 kpc for the metallicity-dependent distances. Since these errors represent the standard deviation of the distance distribution of each globular cluster, the large errors can be largely attributed to the disparity of the distances in each sample. This is likely partly due to contamination by stars not part of the GC and by stars with an incorrectly assigned magnitude. Additionally, the uncertainty in the distance of individual stars is $\sim$ 1 kpc, also contributing to the large variance in distance.

Using the color cut improves the precision by a factor of $\sim$ 1.4. This can be largely attributed to the color cut removing stars that could potentially have an incorrect absolute magnitude, reducing the discrepancy between the distances. However, only three GCs have a reliable distance, based on the size of each GC sample after the cut. This is due to the color cut decreasing the number of sources in each sample significantly. For the other GCs, little improvement is made in the accuracy of the distances. Likely, their distances after the cut are solely based on stars that are not part of the GC or giant stars with incorrect magnitudes that fell just outside the cut. Based on the CMDs a slightly redder color cut could be considered. E.g. the selection of NGC5466 extends past $G - G_{RP} = 0.45$ in the MS turn-off. However, this would lead to an even greater reduction of the selection sample sizes. Increasing the upper limit is also not likely to be advantageous in this case, since very few stars in the CMDs have $G - G_{RP} > 0.715$. Additionally, MS stars corresponding to these colors are generally faint and cover a larger range of $G - G_{RP}$, increasing the error of these distances. In general, using a redder color cut is likely to remove more stars with correctly assigned magnitudes than incorrect ones. Hence, it might often not be worth the trade-off.

Furthermore, the sample size is limited by the limitations and uncertainties of Gaia. GCs are highly dense regions. Consequently, Gaia is not able to resolve individual stars around the center of the GCs, explaining the 'hole' visible in fig. 8. Besides the RPM selected halo sample containing only MS stars, the uncertainty of the parameters of stars in dense regions is relatively higher. Due to quality cuts, for example in the reduced proper motion parameters, the sample size is further limited.

Ultimately this shows that the color cut is effective at removing stars with incorrectly assigned absolute magnitudes However, simultaneously it greatly reduces the sample size and removes stars in the MS that are close to the turn-off. Hence it is not always advantageous to use the color cut, depending on the age and metallicity of the stellar population that is being observed. Based on fig. 12a, the metallicity-dependent distances are more accurate than the photometric distances, especially for more metal-poor populations. However, metallicities are often not available and therefore can't be used in the computation of the distance.

4.2 Phlegethon selection

The sample of Phlegethon members from the RPM selected halo sample is similar to the most recent STREAMFINDER algorithm determined sample for Gaia (E)DR3 provided by R. Ibata. However, we probe at least one magnitude m_g fainter, with a limit around 21 instead of 20. This allows for lower surface brightness structures to be traced. The STREAMFINDER selection does extend to higher values of b for the high l-end of the stream. This might be due to difficulties faced with the first fit to a proper motion overdensity due to high contamination in the sample (fig. 10a). Ibata et al. (2021) shows a parallel structure below the main stream around $l \sim 55°$, this could be part of Phlegethon or another smaller stream with similar properties. The streak above the stream around $l \sim -5°$ is not visible in the STREAMFINDER selected sample or any streams in the surrounding area, implying that it is likely other contamination. The functions in proper motions μ_{RA} and μ_{DEC} against the galactic longitude l (figs. 10c and 10d), can be compared to equivalent plots by Ibata et al. (2018). Their orbital solution is similar to the polynomials that were computed. One notable difference is that the STREAMFINDER determined selection contains more stars with $\mu_l > -25$ mas/yr in the region where $l > 60°$.

One of the downsides of the selection method used, as opposed to the STREAMFINDER using the 6D sample, is that is more difficult to claim with high certainty that sources belong to a particular stellar stream. Another downside is the fact that the RPM selected halo sample only contains MS stars, making it impossible to accurately fit a population model to the color-magnitude distribution and determine the age of the population. Additionally, the RPM sample is kinematically biased. The sample lacks stars with small proper motions due to the selection being focused around stars with high tangential velocities (Viswanathan et al., 2022). Due to the lack of line-of-sight velocities, it is also difficult to accurately determine its true orbit. On the contrary, the RPM selected halo sample is not limited by stars being constricted to an orbit. Consequently, this sample could be more effective at showing interesting features like spurs, especially for low surface brightness stars.

Unlike how most other stream selections in the 5D sample are performed (e.g. (Viswanathan et al., 2022)), namely by using a hand-drawn polygon around the stream, Phlegethon members were selected using a selection of overdensities followed by a sequence of polynomial fits. The advantage of the latter is that it could potentially be automated, simplifying the process and making it applicable to other stellar streams. Nonetheless, this method is still biased. Due to the

large contamination around the overdensities in proper motion, a polynomial could not simply be fit to all sources in the sample. Instead, specific points along the overdensity had to be set by eye or be determined through a Gaussian fit to histograms of bins with a width of $\Delta l = 10°$. The latter is also strongly biased through the boundaries that are set for the Gaussian parameters to secure an accurate fit. This issue is amplified when the overdensity cannot be clearly distinguished from the background. This occurred on the edges of the μ_b fit, shown in fig. 10a. Additionally, the method is biased through the chosen maximum distance stars can be removed from the fitted polynomials. It would be better if Gaussian's were fitted perpendicular to the overdensity along its track, allowing for a better estimate of the true width at different points in the track. However, a general problem with trying to fit Gaussian's in these samples is the non-uniform background.

The distance to Phlegethon was determined to be 3.6 ± 0.6 kpc, which is in agreement with the distance of ~ 3.8 kpc reported in Ibata et al. (2018). The large uncertainty in the found value is likely due to similar issues as the GCs distances, discussed in section 4.1. The same paper also reports a metallicity of $[\mathrm{Fe/H}] = -1.56 \pm 0.04$, which is based on two stars from their selection that crossmatched with the SDSS/Seque survey. The metallicity that was found in this thesis using the 38 cross matches with Pristine is $[\mathrm{Fe/H}] = -1.9 \pm 0.5$, which is in agreement with Ibata et al. (2018). One downside of the distribution of the cross matches with Pristine is that they all reside in the high l tail ($l > 55°$) of the stellar stream, due to Pristine's footprint. Ideally, the cross matches would be distributed throughout the whole stream, increasing the probability that the metallicity reflects the metallicity of the entire stream accurately. It would also us to inspect whether any metallicity variation is present throughout the stream.

4.3 Systematic identification of substructures

Locating and identifying stellar streams by eye in the many plots that were created is a strenuous and tedious process. The six stellar streams that were identified were likely only a fraction of the stellar streams that could be observed in the plots. To effectively find stellar streams, the process needs to be simplified by reducing the number of plots that are created that could potentially contain a stream. Eventually, the desired algorithm would locate and isolate stellar streams automatically.

To maximize the number of streams that can be identified through overdensities, the current method needs to be optimized first. One of the problems of the current method is that the proper motions selected for a single plot are within a 5 mas/yr interval. Many stellar streams have proper motions that cover a much larger range. E.g. Phlegethon for which $-20 < \mu_b < 15$ mas/yr and $-40 < \mu_l < -20$ mas/yr. By only probing a small part of the proper motions, the overdensity will not appear as a stream-like structure. Likewise, the density deviation from the background will be smaller, making it less apparent. One could loop over a larger combination of proper motion intervals to limit the problem. However, that would create a significantly larger number of plots, making it an even more lengthy and strenuous process to identify streams. As a solution, more conditions should be put in place that decides whether a plot is created or not. Currently, only one condition is in place, namely, the number of sources in the selection must exceed 2000.
If the proper motions of a stream deviate enough from the background, a peak in proper motion should be present. By selecting stars around the peak, part of the contamination can already be removed. This method would be most effective if the proper motion is approximately constant in at least one direction, as this would result in a more prominent peak. Ideally, one would use

a map of the background proper motions, allowing streams to be observed through the residual. However, the proper motions of stellar streams do not always differ significantly from their background in the sky. Another possibility could be to use the correlation between densities along two directions. For a stellar stream, it is likely there will be a positive correlation, due to the density in both directions increasing at these points. One of the issues with this method is that streams are usually thin structures, which are difficult to isolate if the stream is not aligned with one of the two directions.

Additionally, some substructures appear due to the scanning pattern of Gaia. The scanning pattern is created, due to some areas being scanned more frequently than others, allowing more and fainter stars to be detected in these regions. Gaia's scanning pattern can be observed in Riello, M. et al. (2021). Examples of overdensities due to the scanning pattern were shown in section 3.2.2. This problem could be reduced by using a model of the scanning pattern, which can be scaled and subtracted from density plots. Another method could be to automatically cross-reference any overdensities that are found with the scanning pattern and let them be disregarded in case of a match. Another concern is the non-uniform background. Towards the galactic plane, the density of sources increases substantially, partly due to the contamination of stars of the thick disk. This increases the complexity of isolating overdensities. Selected overdensities should be divided by their shape. Overdensities corresponding to the background will be larger in size, follow a large-scale gradient in density, and will not have a stream-like shape.

Once a better selection process is in place, the boundary parameters can be relaxed to allow for nearby structures, and structures that lie closer to the galactic plane to be detected as well. The example of Hríd and Gjöll, shown in section 3.2, shows that it is feasible to observe these types of stellar streams.

Laporte et al. (2021) shows it is possible to detect substructures using peaks in proper motion space in the outer disk. They spatially bin the data into bins of 0.5° by 0.5° and apply a so-called "shrinking sphere algorithm" to find proper motion peaks in each pixel. By mapping these peaks, several new substructures were identified. A similar method could be applied to the galactic halo. However, likely larger bins have to be used due to the lower density of the halo.

In the end, one should have an algorithm that can locate stellar streams using the RPM selected halo sample from Gaia. This sample consists of around 47 million halo stars, as opposed to the 7.2 million stars in the 6D sample of Gaia EDR3 (Gaia Collaboration et al., 2018). Additionally, the RPM selected halo sample contains photometric distance with an uncertainty of $\approx 7\%$ (Viswanathan et al., 2022), which is more precise and accurate than the distances that Gaia parallaxes provide. Under the assumption that the variance in distance within a stellar stream is negligible compared to the heliocentric distance to the stream, slices in distance spanning interval 2-4 times the uncertainty of the photometric distance, allows streams to more effectively be isolated.

5 Conclusion

In this thesis, we explored the detection of stellar streams in the Milky Way halo. A stellar stream is created when a dwarf galaxy or globular cluster merges with the Milky Way, causing it to be torn apart by tidal forces. Identifying stellar streams can provide information about their progenitors and the history of the Galaxy.
Digital sky surveys like SDSS and DES have enabled the discovery of many streams in the outer halo. New data that is provided by Gaia, has allowed for new streams to be discovered and for known streams to be studied in more detail. The STREAMFINDER algorithm has played an instrumental role in the study of many stellar streams. The algorithm has automated the search for stellar streams using the 6D data provided by Gaia. Despite being very effective, the 6D sample only consists of around 7.2 million stars.
The data analysis done in this thesis has used the RPM selected halo sample provided by Viswanathan et al. (2022). This sample of around 48 million stars, contains photometric distances that are more reliable than the distances provided by the Gaia parallaxes.

By computing the distances of seven globular clusters with known distances and metallicities, the reliability of the photometric distances compared to distances computed using the metallicity was tested. It was also concluded that the color cut ($0.45 > G - G_{RP} > 0.715$) is successful in removing stars with incorrectly assigned absolute magnitudes. However, it also greatly decreased the sample size of each GCs. Consequently, for only three GCs accurate distances were computed. After the color cut, there is little difference in accuracy between the photometric and metallicity-dependent distances.

The next part explores the first step of systematically and eventually automatically detecting stellar streams. By probing the RPM selected halo sample at different proper motion and space intervals a series of scatter plots were created through which several stellar streams were identified by eye. Potential future improvements need to put more conditions in place that limit the number of plots that potentially contain observable streams. This would also enable the boundary parameters to be relaxed to explore are a larger range of proper motions, distances, and space.

Finally, a selection of potential members of the stellar stream Phlegethon was created using a sequence of alternating polynomial fits to the overdensity in proper motions μ_b, μ_l, μ_{RA} and μ_{DEC} and the stream-track. Stars within a given width around the polynomial were selected for the next fit in the sequence. The final sample consists of 575 stars with a mean photometric distance of 3.6 ± 0.6 kpc. Using a crossmatch with Pristine, a metallicity of $[\mathrm{Fe/H}] = -1.9 \pm 0.5$ was established. Phlegethon spans between $-15° < l < 70°$ and $-50° < b < -25°$ and is strongly retrograde. This is in agreement with what has previously been reported on Phlegethon. A crossmatch with the STREAMFINDER selection of Phlegethon by R. Ibata shows that this method can successfully isolate stars corresponding to the stream. The main difference is that the STREAMFINDER selection extends to higher values of b for the high l-end of the stream. However, our selection shows more extending off-stream features. The main shortcoming of the used method is that is biased by the widths of the selections that are set and the selected points on the overdensities in proper motion. The selection of Phlegethon candidates provides a proof of concept that stellar streams can be selected through a simple method that can potentially be automated, bringing us one step closer to creating an algorithm that can systematically and automatically detect stellar streams with the 5D data sample.

6 Acknowledgements

I am extremely grateful to my supervisors Else Starkenburg and Akshara Viswanathan for their support during the project. They were always available for questions and feedback and provided helpful and important insights. Akshara has also been incredibly helpful with more specific coding-related questions and accessing the data. She also performed the crossmatch with Pristine to provide me with metallicities for Phlegethon.

I would like to thank Rodrigo Ibata for the Phlegethon candidates from STREAMFINDER. This allowed me to accurately check the quality of my selection and compare the features of the stream in more detail.

This work has made use of data from the European Space Agency (ESA) mission *Gaia* (`https://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, `https://www.cosmos.esa.int/web/gaia/dpac/consortium`). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement.

References

Barnes J. E., 1992, ApJ, 393, 484

Belokurov V., et al., 2007, ApJ, 658, 337

Bensby T., Feltzing S., Lundström I., 2003, A&A, 410, 527

Bhardwaj A., 2018, Cepheid and RR Lyrae Variables as Standard Candles and What Else?, doi:10.48550/ARXIV.1806.03021, https://arxiv.org/abs/1806.03021

Bland-Hawthorn J., Gerhard O., 2016, ARA&A, 54, 529

Bonaca A., Hogg D. W., Price-Whelan A. M., Conroy C., 2019, ApJ, 880, 38

Breddels M. A., Veljanoski J., 2018, A&A, 618, A13

De Silva G. M., Sneden C., Paulson D. B., Asplund M., Bland-Hawthorn J., Bessell M. S., Freeman K. C., 2006, AJ, 131, 455

ESA 2020, Gaia Mission Science Performance - Gaia - Cosmos, https://www.cosmos.esa.int/web/gaia/science-performance#astrometric%20performance

Fattahi A., et al., 2020, MNRAS, 497, 4459

Gaia Collaboration et al., 2018, A&A, 616, A1

Gaia Collaboration et al., 2021, A&A, 650, C3

Gaia Collaboration Vallenari, A. Brown, A.G.A. Prusti, T. et al. 2022, A&A

Girelli G., Pozzetti L., Bolzonella M., Giocoli C., Marulli F., Baldi M., 2020, A&A, 634, A135

Grillmair C. J., Carlin J. L., 2016, in Newberg H. J., Carlin J. L., eds, Astrophysics and Space Science Library Vol. 420, Tidal Streams in the Local Group and Beyond. p. 87 (arXiv:1603.08936), doi:10.1007/978-3-319-19336-6_4

Gómez F. A., Helmi A., Cooper A. P., Frenk C. S., Navarro J. F., White S. D. M., 2013, Monthly Notices of the Royal Astronomical Society, 436, 3602

Harris W. E., 1996, AJ, 112, 1487

Haywood M., Di Matteo P., Lehnert M. D., Snaith O., Khoperskov S., Gómez A., 2018, ApJ, 863, 113

Helmi A., 2020, ARA&A, 58, 205

Helmi A., White S. D. M., 1999, MNRAS, 307, 495

Helmi A., White S. D. M., de Zeeuw P. T., Zhao H., 1999, Nature, 402, 53

Helmi A., Babusiaux C., Koppelman H. H., Massari D., Veljanoski J., Brown A. G. A., 2018, Nature, 563, 85

Ibata R. A., Gilmore G., Irwin M. J., 1994, Nature, 370, 194

Ibata R. A., Malhan K., Martin N. F., Starkenburg E., 2018, ApJ, 865, 85

Ibata R., et al., 2021, ApJ, 914, 123

Ivezić Ž., et al., 2000, AJ, 120, 963

Johnston K. V., 1998, ApJ, 495, 297

Johnston K. V., 2016, in Newberg H. J., Carlin J. L., eds, Astrophysics and Space Science Library Vol. 420, Tidal Streams in the Local Group and Beyond. p. 141 (arXiv:1603.06601), doi:10.1007/978-3-319-19336-6_6

Jurić M., et al., 2008, ApJ, 673, 864

Koppelman H. H., Helmi A., 2021, A&A, 645, A69

Koppelman H., Helmi A., Veljanoski J., 2018, ApJL, 860, L11

Koppelman H. H., Helmi A., Massari D., Roelenga S., Bastian U., 2019, A&A, 625, A5

Laporte C. F. P., Koposov S. E., Belokurov V., 2021, Monthly Notices of the Royal Astronomical Society: Letters, 510, L13

Lövdal S., Ruiz-Lara T., Koppelman H., Matsuno T., Dodd E., Helmi A., 2022, Astronomy &amp$$ Astrophysics

Malhan K., Ibata R. A., 2018, MNRAS, 477, 4063

Malhan K., Ibata R. A., Martin N. F., 2018, MNRAS, 481, 3442

Martin N. F., et al., 2022, The Pristine survey – XVI. The metallicity of 26 stellar streams around the Milky Way detected with the STREAMFINDER in Gaia EDR3, doi:10.48550/ARXIV.2201.01310, https://arxiv.org/abs/2201.01310

Martinez-Valpuesta I., Gerhard O., 2013, ApJL, 766, L3

Massari D., Koppelman H. H., Helmi A., 2019, A&A, 630, L4

McConnachie A. W., 2012, AJ, 144, 4

McMillan P. J., 2017, MNRAS, 465, 76

Mo H. J., Mao S., White S. D. M., 1998, MNRAS, 295, 319

Newberg H. J., Willett B. A., Yanny B., Xu Y., 2010, ApJ, 711, 32

Nibauer J., Belokurov V., Cranmer M., Goodman J., Ho S., 2022, arXiv e-prints, p. arXiv:2205.11767

Portail M., Gerhard O., Wegg C., Ness M., 2017, MNRAS, 465, 1621

Quinn P. J., 1984, ApJ, 279, 596

Renaud F., Agertz O., Gieles M., 2017, MNRAS, 465, 3622

Riello, M. et al., 2021, A&A, 649, A3

Rodriguez-Gomez V., et al., 2016, MNRAS, 458, 2371

Ruiz-Lara T., Matsuno T., Sofie Lövdal S., Helmi A., Dodd E., Koppelman H. H., 2022, arXiv e-prints, p. arXiv:2201.02405

Schönrich R., Binney J., Dehnen W., 2010, MNRAS, 403, 1829

Shipp N., et al., 2018, ApJ, 862, 114

Tolstoy E., Venn K. A., Shetrone M., Primas F., Hill V., Kaufer A., Szeifert T., 2003, AJ, 125, 707

Tolstoy E., Hill V., Tosi M., 2009, ARA&A, 47, 371

Tremonti C. A., et al., 2004, ApJ, 613, 898

Valenti E., et al., 2016, A&A, 587, L6

Virtanen P., et al., 2020, Nature Methods, 17, 261

Viswanathan A., Starkenburg E., Koppelman H. H., Helmi A., 2022, preprint

Wang J., et al., 2011, MNRAS, 413, 1373

Wegg C., Gerhard O., Portail M., 2015, MNRAS, 450, 4050

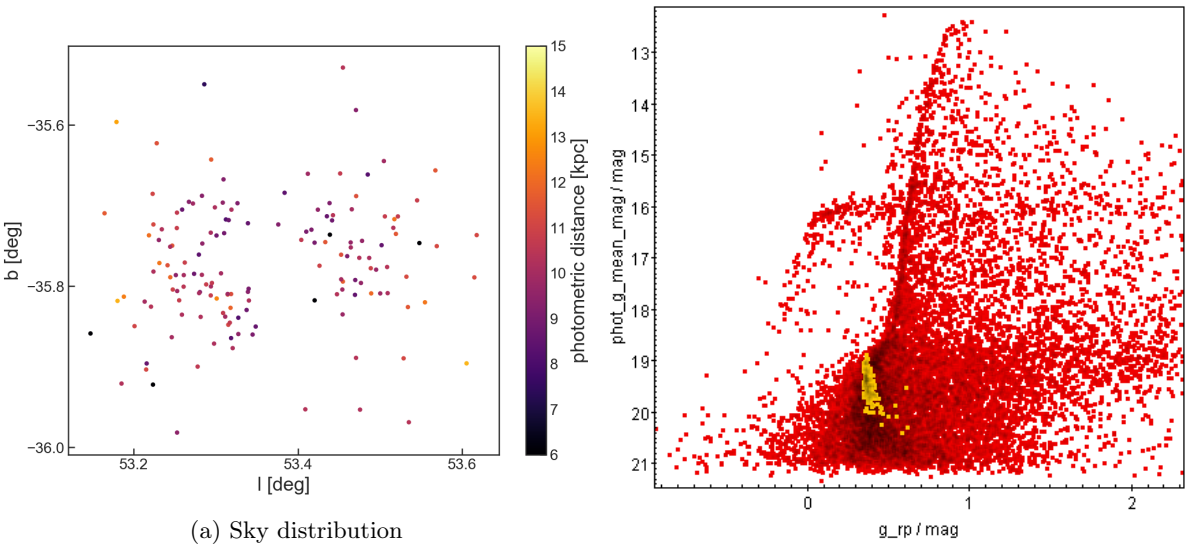
White S. D. M., Rees M. J., 1978, MNRAS, 183, 341

Yanny B., et al., 2000, ApJ, 540, 825

de Boer T. J. L., Belokurov V., Koposov S., 2015, MNRAS, 451, 3489

A Globular Clusters: Sky distribution and CMD

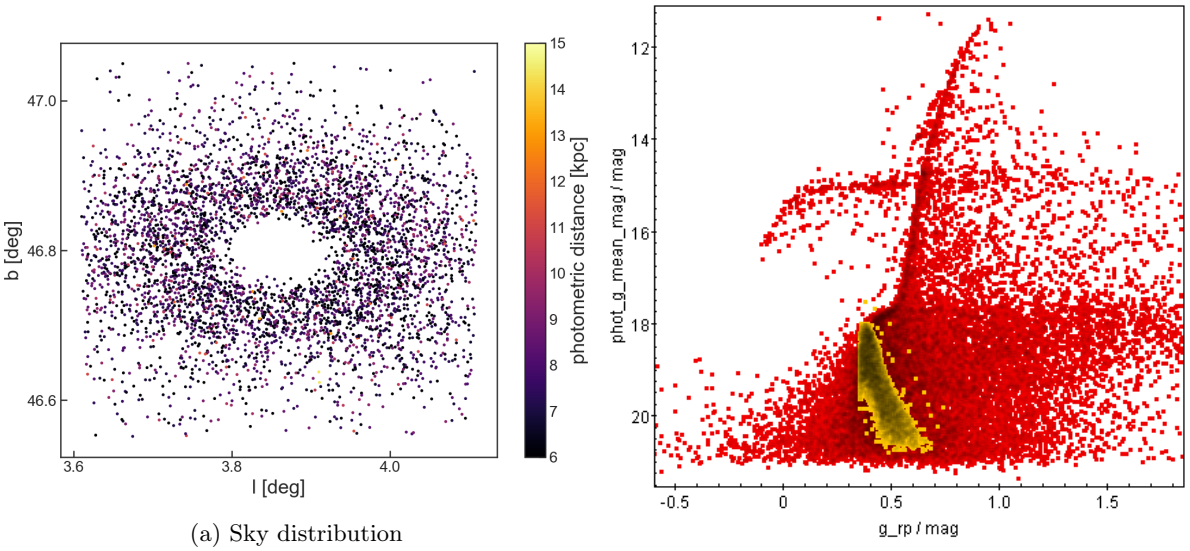
For seven globular clusters a CMD is created. The selection of each GC in the RPM selected halo sample without the color-cut (yellow) is layed on top of the selection with the full Gaia EDR3 sample (red). The latter was created by selecting all stars within 0.15° of the GC's center. This is a rough selection, consequently it will include more contamination. The CMD plot will show whether the stars in the selection belong to the MS or the MS turn-off.



(a) Sky distribution

(b) CMD

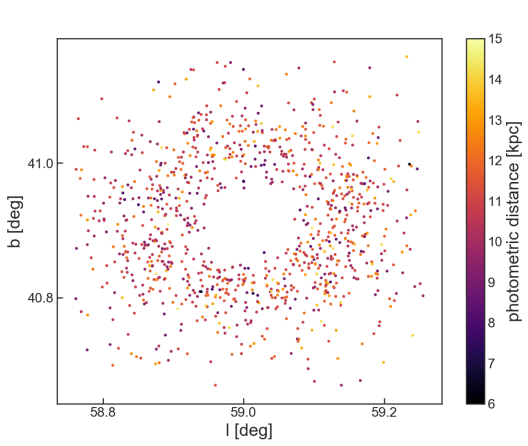
Figure 28: M2, NGC7089



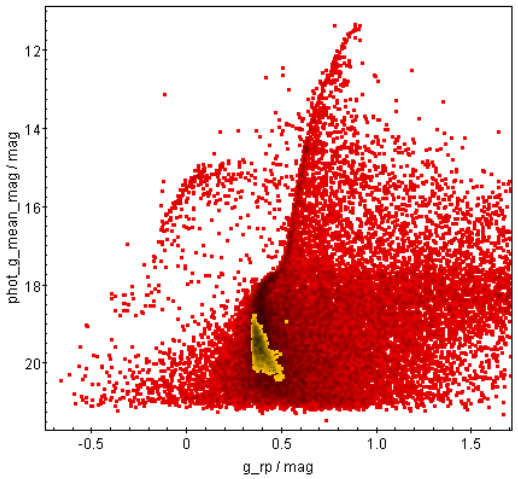
(a) Sky distribution

(b) CMD

Figure 29: M5, NGC5904

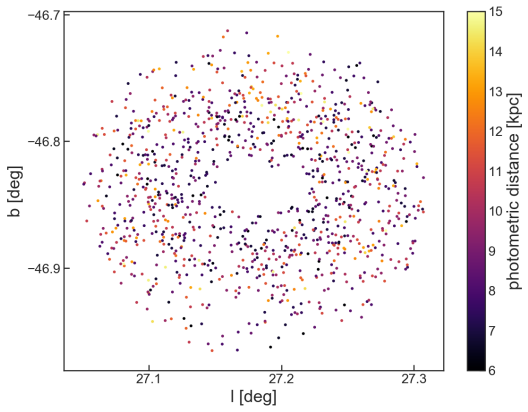


(a) Sky distribution

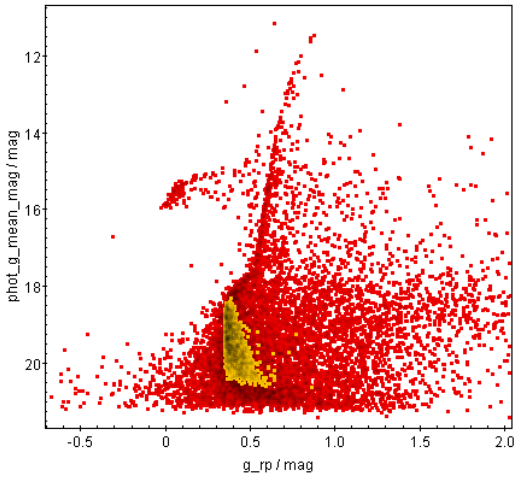


(b) CMD

Figure 30: M13, NGC6205

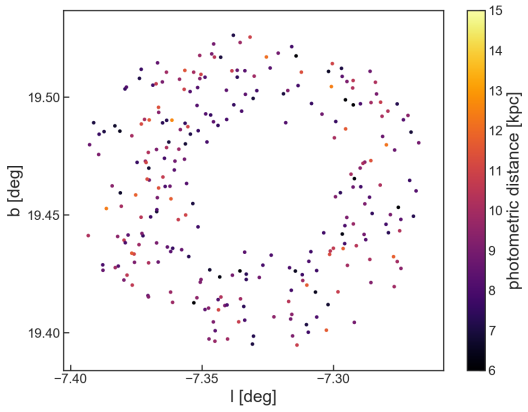


(a) Sky distribution

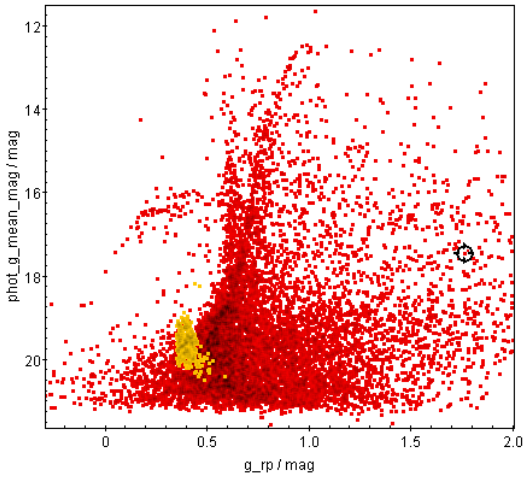


(b) CMD

Figure 31: M30, NGC7099

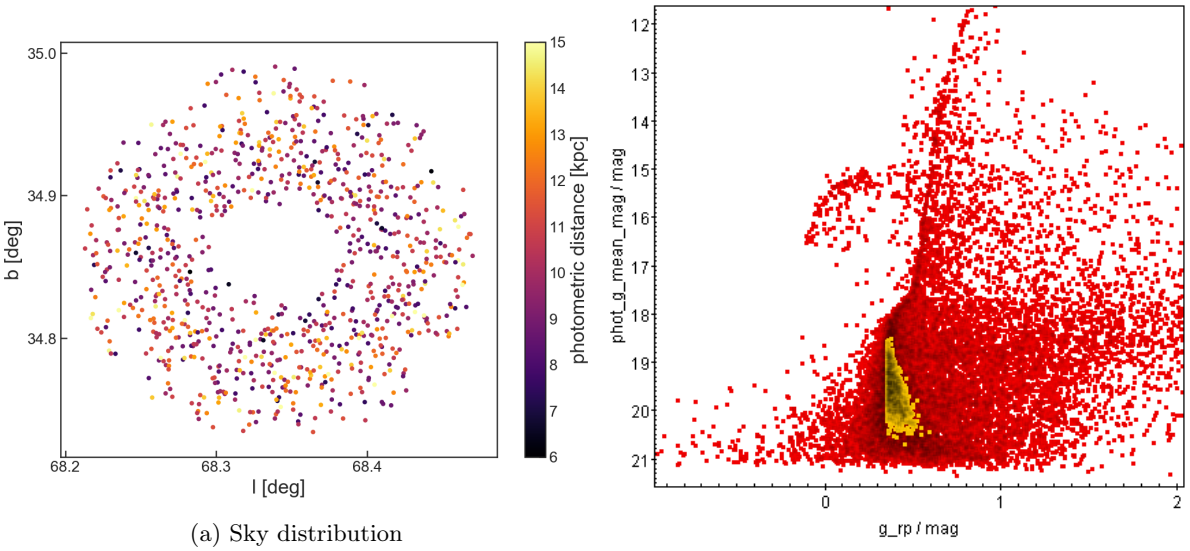


(a) Sky distribution



(b) CMD

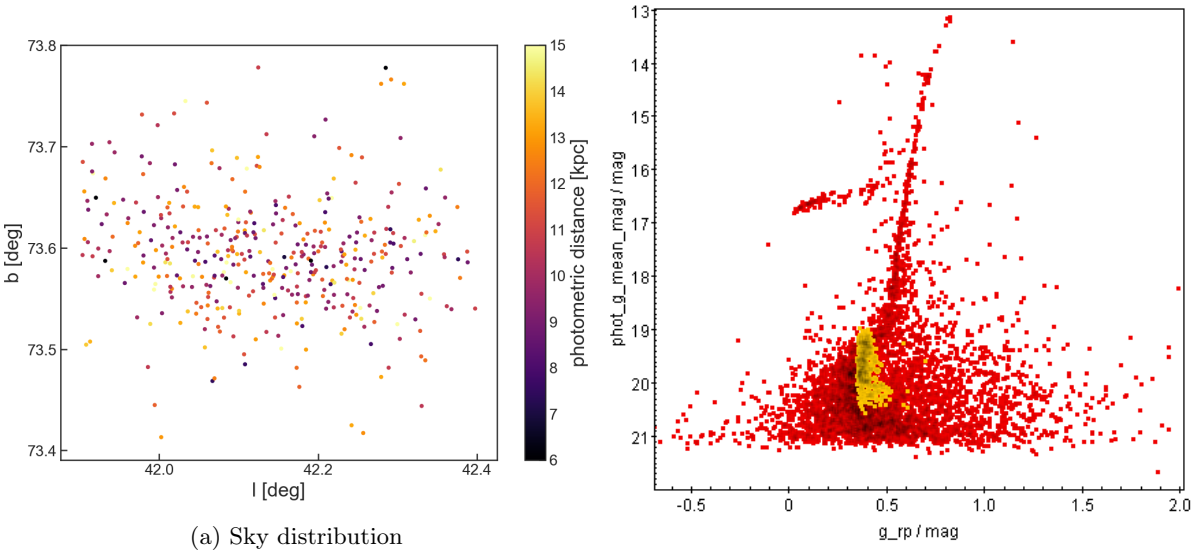
Figure 32: M80, NGC6093



(a) Sky distribution

(b) CMD

Figure 33: M92, NGC6341



(a) Sky distribution

(b) CMD

Figure 34: NGC5466

B Python Code

```
# ## Import modules

# In[1]:

import vaex
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.pyplot import figure, show
from scipy.optimize import curve_fit
from astropy.convolution import Gaussian2DKernel, convolve


# ## Import data

# In[2]:

df = vaex.open('/net/gaia2/data/users/viswanathan/anna-esselink/RPM-full-catalog-edr3.hdf5')
df


# In[3]:

df["l_adj"] = df.l

#quality cuts
Select_qual = (df.log10_Hg_over_error>1.75) &((df.g_rp_corrected>0.45)& \
                                               (df.g_rp_corrected<0.715))
Select_qual2 = (df.log10_Hg_over_error>1.75)

# adjusting range of l from 0-360 to -180-180
Select = (df.l_adj>=180)
df['l_adj'] = df.func.where(Select, (df.l_adj-360), df.l_adj)


# In[4]:

# Calculate and add longitudinal and latitudinal velocities to df

#solar motion in km/s
U_sol = 11.1
V_sol = 12.24
W_sol = 7.25
v_LSR = 232.8  #motion of the local standard of rest

df.add_virtual_column('v_l_sol', (-U_sol*np.sin(df.l*np.pi/180) +    \
                                  (V_sol + v_LSR)*np.cos(df.l*np.pi/180)))
df.add_virtual_column('v_b_sol', (W_sol*np.cos(df.b*np.pi/180) - np.sin(df.b*np.pi/180)   \
          *(U_sol*np.cos(df.l*np.pi/180) + (V_sol + v_LSR)*np.sin(df.l*np.pi/180))))
df.add_virtual_column('v_l','4.74057*pm_l_apex*phot_dist')
df.add_virtual_column('v_b','4.74057*pm_b_apex*phot_dist')

# proper motions not corrected for solar motion
df.add_virtual_columns_proper_motion_eq2gal(long_in="ra", lat_in="dec",
            pm_long="pmra", pm_lat="pmdec", pm_long_out="pml", pm_lat_out="pmb",
                                    radians=False,propagate_uncertainties=True)

df
```

B.1 Globular cluster distance

```
# ## Globular clusters - Distances
#

# In[4]:


# run this notebook to calculate magnitude using the metallicity and g_rp with the msfehfit() function
get_ipython().run_line_magic('run', '/net/gaia2/data/users/viswanathan/anna-esselink/ \
                                                    metallicity-dependent-distances.ipynb')

# In[5]:

# access GC catalog
path = '/net/gaia2/data/users/koppelman/files-from-arend/koppelman/Research/ \
                                                        DR2_6D/GC_AGE_METALLICITY/'
gc= vaex.open(path + 'GC-cat-full.v4.hdf5')
Names = gc.Name.values

# extract metallcity and distance from the GC catalog
Met = np.array([])
Dist = np.array([])
GC_names = [b'NGC7089', b'NGC5904', b'NGC6205', b'NGC6402', b'NGC7099', b'NGC6093', \
                                                    b'NGC6341', b'NGC5466']
for name in GC_names:
    i = int(np.where(Names==name)[0])
    Met = np.append(gc.met.values[i], Met)
    Dist  = np.append((gc.x.values[i]**2 + gc.y.values[i]**2 + gc.z.values[i]**2)**0.5, Dist)
#     print(Met)

Met = np.flip(Met)
Dist = np.flip(Dist)


# In[6]:


def Gaussian(x,a,b,c,d):
    """Returns a gaussian function with independent variable x, and parameters a,b,c,d"""
    return a*np.exp(-(x-b)**2/(2*c*c)) + d

def glob_cluster_dist(l,b, color_cut=False, plot_hist=False, plot_scatter=False,
                                                    parr = False, CMD = True, i=0):
    """Based the galactic coordinates l and b (in degrees) of the GC's center, this
    function isolates stars belonging to the GC through its proper motion and
    density distribution. i is the index of the GC in the list:
    [NGC7089, NGC5904, NGC6205, NGC6402, NGC7099, NGC6093, NGC6341, NGC5466]

     if  color_cut=True      -   The color cut 0.45 < g_rp < 0.715 is applied
     if  plot_hist=True      -   plot a histogram of the photometric and parallax distance
     if  plot_scatter=True   -   plot a scatterplot of the final selection of stars
     if  parr=True           -   the output will include the mean parallax distance
     if  CMD=True            -   plot of the CMD of bp_g vs. gmag_corrected
    """
    Select_qual2 = (df.log10_Hg_over_error>1.75)

    hw = 0.25    #half width of globular cluster "area" [deg]
    select_sky = ((df.l_adj>l-hw) & (df.l_adj< l+hw) & (df.b>b-hw) & (df.b<b+hw))

    Select =  (Select_qual2 & select_sky)   #& select_dist

    #background removement through selection of proper motion peak
    counts_pm = df.count(binby="pm_l_apex", limits=[-20, 20], selection=Select, shape=100)
    pm_bins = np.linspace(-20,20,100)
    pm_peak = pm_bins[np.argmax(counts_pm)]

    select_pm = (df.pm_l_apex> pm_peak-3) & (df.pm_l_apex < pm_peak+3)
    Std  = df.pm_l_apex.std(selection=(Select & select_pm))
```

```
    select_pm2 = (df.pm_l_apex> pm_peak-Std*1.3) & (df.pm_l_apex < pm_peak+Std*1.3)
    Select2 =  (Select_qual2 & select_sky  & select_pm2)   #& select_dist

    # parameters converted to array for convenience
    l_adj_arr = df[Select2].l_adj.values
    b_arr = df[Select2].b.values
    pm_arr = df[Select2].pm_l_apex.values
    dist_arr = df[Select2].phot_dist.values

    r = ((b_arr-b)**2 + (l_adj_arr-l)**2)**0.5   #distance from centre

    #finer cut around GC
    counts_l = df.count(binby="l_adj", limits=[l-hw,l+hw], selection=Select2, shape=50)
    l_bins = np.linspace(l-hw,l+hw,50)
    popt,pcov = curve_fit(Gaussian, l_bins, counts_l, p0=([30,l,0.1,10]))
    a_fit, b_fit, c_fit, d_fit = popt
#     print(a_fit, b_fit, c_fit, d_fit)
    R = 1.0 *2.4*c_fit/2  #circular (radius=R) cut at 1.*FWHM/2

    select_sky2 = np.where((r<R)&(r>0*R))


    print(len(l_adj_arr[select_sky2]))

    if plot_scatter==True:

        plt.figure(figsize = (9,7))
        plt.scatter(l_adj_arr[select_sky2], b_arr[select_sky2], s=2, alpha=1,
                    c=dist_arr[select_sky2], cmap='inferno', vmin=6, vmax=15)
        plt.colorbar()
        plt.xlabel('l [deg]', fontsize =14)
        plt.ylabel('b [deg]', fontsize=14)
        plt.xticks(fontsize=14)
        plt.yticks(fontsize=14)
        plt.show()

    if plot_hist==True:
        df.viz.histogram(df.distance, limits=[-10,20], selection=Select2, shape=100)
        df.viz.histogram(df.phot_dist, limits=[-10,20], selection=Select2, shape=100)

    if CMD==True:
        bp_g = df[Select2].g_rp_corrected.values
        g_mag = df[Select2].gmag_corrected.values
        pm_arr = df[Select2].pm_l_apex.values


        plt.figure(figsize = (8,8))
        plt.scatter(bp_g[select_sky2], g_mag[select_sky2], s=2, alpha=0.5,
                    c=pm_arr[select_sky2], cmap='inferno', vmin=-10, vmax=10)

        plt.colorbar()
        plt.gca().invert_yaxis()
        plt.xlabel('bp-g', fontsize =14)
        plt.ylabel('g', fontsize=14)
        plt.xticks(fontsize=14)
        plt.yticks(fontsize=14)
        plt.show()

    #selection stars within FWHM of the centre
    Select3 =  (Select_qual2  & select_pm2 & select_sky& ((((df.b-b)**2 +
               (df.l_adj-l)**2)**0.5 )<R) & ((((df.b-b)**2 + (df.l_adj-l)**2)**0.5 )>0*R))

    if color_cut==True:
        Select3 =  Select3 & ((df.g_rp_corrected>0.45)&(df.g_rp_corrected<0.715))

    N= df[Select3].length()

    # photometric distance
    Dist_GC = df.mean("phot_dist", selection=Select3)
```

```
137     Std_GC = df.std("phot_dist", selection=Select3)
138     print(df.mean("phot_dist_uncertainty", selection=Select3))
139 
140     if parr==True:   #parallax distances
141         Dist_GC_parr = df.mean("distance", selection=Select3)
142         Std_GC_parr = df.std("distance", selection=Select3)
143         return Dist_GC, Std_GC, Dist_GC_parr, Std_GC_parr
144 
145     # metallicity dependent distance calculation
146     g_rp_corrected_arr = df[Select3].g_rp_corrected.values
147     gmag_corrected_arr = df[Select3].gmag_corrected.values
148     Gmag  = msfehfit((Met[i], g_rp_corrected_arr))
149     M_EG_arr= df[Select3].M_EG.values
150 
151     dist_arr = (10**(1+(gmag_corrected_arr-M_EG_arr-Gmag)/5))/1000
152     dist = [np.nanmean(dist_arr), np.nanstd(dist_arr)]
153 
154     return Dist_GC, Std_GC, dist
155 
156 # Example for M92, NGC6341
157 Dist_6341, Std_6341, dist6341 = glob_cluster_dist(68.34, 34.86, plot_hist=False,
158                                     plot_scatter=True, parr=False, CMD=True, i=6)
159 print(f"M92: d_phot = {Dist_6341}   {Std_6341} kpc" )
```

B.2 Systematic search

```
# ## Plots pm

# In[4]:

select_dist = (df.phot_dist>8)& (df.phot_dist<25)  # distance slice (in kpc)
Select_qual2 = (df.log10_Hg_over_error>1.75)

#lower limits proper motion bins
pm_bin_size = 5
pm_l_low = np.arange(-10,10,pm_bin_size)
pm_b_low = np.arange(-10,10,pm_bin_size)

#lower limits coordinates bins
coord_bin_sizel = 60
coord_bin_sizeb = 60
l_low = np.flip(np.arange(-120,240, coord_bin_sizel))
b_low = np.array([-90, 30])

# loop over spatial coordinates and proper motions in l and b
for lim_l in l_low:
    for lim_b in b_low:
        for lim_pml in pm_l_low:
            for lim_pmb in pm_b_low:

                select_pm = (df.pm_l_apex>lim_pml) & (df.pm_l_apex < lim_pml+pm_bin_size )
                          & (df.pm_b_apex>lim_pmb) & (df.pm_b_apex < lim_pmb+pm_bin_size )
                select_sky = (df.l_adj>(lim_l - coord_bin_sizel)) & (df.l_adj< lim_l)
                                  & (df.b>lim_b) & (df.b< (lim_b + coord_bin_sizeb))

                Select = select_pm & Select_qual2 & select_dist & select_sky

                N = df[Select].length()    #number of stars

                # for all selections with more than 2000 stars a scatter plot is created
                if (N > 2e3):
                    print(f"{lim_pml} < pm_l < {lim_pml+pm_bin_size}, {lim_pmb} \
____________________________________________< pm_b < {lim_pmb+pm_bin_size}, {N}")

                    #scatter plot
                    l_adj_arr = df[Select].l_adj.values
                    b_arr = df[Select].b.values
                    pm_arr = df[Select].pm_l_apex.values

                    plt.figure(figsize = (9,7))
                    plt.scatter(l_adj_arr, b_arr, s=3, alpha=0.3, c=pm_arr, cmap='inferno')
                    plt.colorbar()
                    plt.xlim([lim_l, lim_l-coord_bin_sizel])
                    plt.ylim([lim_b, (lim_b+coord_bin_sizeb)])
                    plt.xlabel('l [deg]')
                    plt.ylabel('b [deg]')
                    plt.show()
```

B.3 Phlegethon star selection

```
# ## Phlegeton selection

# In[5]:

def Norm_Data(data):
    """Normalises the dataset"""
    return (data - np.min(data))/(np.max(data) - np.min(data))

def Gaussian(x,a,b,c,d):
    """Returns a gaussian function with independent variable x, and parameters a,b,c,d"""
    return a*np.exp(-(x-b)**2/(2*c*c)) + d

# polynomical functions
def linear(l, a, b):
    return a*l + b

def quadratic(x, a, b, c):
    return a*x**2 + b*x +c

def cubic(x, a, b, c, d):
    return a*x**3 + b*x**2 + c*x + d


# In[6]:

select_qual = (df.log10_Hg_over_error>1.75)
pixel_size = 1  # 0.5 degree per pixel

#phlegethon
# initial rough selection in area, distance and proper motion
region_sky = [[-20, 70],[-50,-35]]
select_dist = (df.phot_dist>2.5)& (df.phot_dist<4.5)
select_pm = (df.pm_l_apex>-40) & (df.pm_l_apex < -20) & (df.pm_b_apex>-30) & \
                                                        (df.pm_b_apex < 5)
select_sky = ((df.l_adj> region_sky[0][0]) & (df.l_adj< region_sky[0][1]) & \
                            (df.b> region_sky[1][0]) & (df.b< region_sky[1][1]))

Select = (select_pm & select_qual & select_dist & select_sky)

# relevant parameters are converted to arrays
l_adj_arr = df[Select].l_adj.values
b_arr = df[Select].b.values
pm_arr = df[Select].pm_l_apex.values

#density plot
Shape = (int((region_sky[0][1]-region_sky[0][0])/pixel_size), \
                    int((region_sky[1][1]-region_sky[1][0])/pixel_size))  # 2 pixels per degree

# 2d histogram is created for a given binning (shape)
lbcounts = df.count(binby=[df.l_adj, df.b], selection=Select, shape=Shape, limits=region_sky)
kernel = Gaussian2DKernel(x_stddev=1.5)  #gaussian smoothing
counts_smooth = Norm_Data(convolve(lbcounts, kernel))  #normalisation

# extent of plot in galactic coordinates
Ext = [100, 220, 40, 65]
Ext = [region_sky[0][0], region_sky[0][1], region_sky[1][0], region_sky[1][1]]

# density plot of 'rough selection'
plt.figure(figsize = (10,2))
# plt.imshow(lbcounts.T, origin='lower', extent=Ext, cmap='inferno')
im = plt.imshow(counts_smooth.T, origin='lower', extent=Ext, cmap='inferno')
cbar = plt.colorbar(im)
plt.gca().invert_xaxis()
cbar.set_label('counts', rotation=90)
plt.xlabel('l [deg]')
plt.ylabel('b [deg]')
plt.show()
```

```
68
69  # same plot but a contour is drawn around the high density areas
70  plt.figure(figsize = (10,2))
71  im = plt.imshow(counts_smooth.T, origin='lower', extent=Ext, cmap='inferno')
72  plt.contour(counts_smooth.T,levels=[0.6], extent=Ext, colors='yellow')
73  cbar = plt.colorbar(im)
74  plt.gca().invert_xaxis()
75  cbar.set_label('counts', rotation=90)
76  plt.xlabel('l [deg]')
77  plt.ylabel('b [deg]')
78  plt.show()
79
80
81  # In[7]:
82
83
84  l_coords = np.arange(region_sky[0][0], region_sky[0][1], pixel_size)
85  b_coords= np.arange(region_sky[1][0], region_sky[1][1], pixel_size)
86
87  l_grid, b_grid = np.meshgrid(l_coords, b_coords)
88  l_grid = l_grid.flatten() ; b_grid = b_grid.flatten()
89
90  # all pixels with a higher density (within the contour) are selected
91  ind = np.argwhere(counts_smooth.T.flatten()>0.6)
92
93
94  select_dist = (df.phot_dist>2.5)& (df.phot_dist<4.5)
95  select_pm = (df.pm_l_apex>-40) & (df.pm_l_apex < -19) & (df.pm_b_apex>-30) & (df.pm_b_apex < 10)
96  select_mag2 = df.gmag_corrected<19.5
97
98  #select all stars within contour
99  select_rest = Select_qual2 & select_dist & select_pm  & select_mag2  #& select_l
100 l_adj_arr = df[select_rest].l_adj.values
101 b_arr = df[select_rest].b.values
102 select1 = np.where(df.l_adj>1000, False, False)
103
104 # loop over every pixel with high density and label stars with galactic coordinates
105 # corresponding to that pixel
106 for i in ind[:]:
107     select = np.where(((l_adj_arr> l_grid[i]) & (l_adj_arr< l_grid[i] + pixel_size) & \
108                        (b_arr> b_grid[i]) & (b_arr< b_grid[i] + pixel_size)), True, False)
109     select1 = select1 | select
110
111 # all stars in the high density area are put into a selection
112 phleg_select1 = {"l_adj": np.array([]), "b": np.array([]), "pm_l_apex": np.array([]),
113     "pm_b_apex": np.array([]), "phot_dist": np.array([]), "gmag_corrected": np.array([])}
114 for key in phleg_select1:
115     phleg_select1[key] = eval(f"df[select_rest].{key}.values[select1]")
116
117
118
119 # In[8]:
120
121
122 # fit 2nd deg. polynomial to stream track of thecurrent selected sample
123 popt,pcov = curve_fit(quadratic, phleg_select1["l_adj"], phleg_select1["b"])
124 a_fit, b_fit, c_fit = popt
125 print('orbit fit 1:')
126 print(popt, np.sqrt(np.diag(pcov)))
127
128 # model of dictionary with relevant parameters
129 phleg_select_empty = {"l_adj": np.array([]), "b": np.array([]), "pm_l_apex": np.array([]),
130     "pm_b_apex": np.array([]), "pml": np.array([]), "pmb": np.array([]), "pmra": np.array([]),
131     "pmdec": np.array([]), "pmra_error": np.array([]), "pmdec_error": np.array([]),
132     "pml_uncertainty": np.array([]), "pmb_uncertainty":np.array([]), "phot_dist": np.array([]),
133     "gmag_corrected": np.array([]), "source_id": np.array([])}
134
135
136 #select all stars within 8 deg in b direction of orbit fit
```

```
width = 8
select2 = np.where(((b_arr > quadratic(l_adj_arr, a_fit, b_fit, c_fit) - width) &  \
           (b_arr < quadratic(l_adj_arr, a_fit, b_fit, c_fit) + width)), True, False)
phleg_select2 = phleg_select_empty.copy()
for key in phleg_select2:
    phleg_select2[key] = eval(f"df[select_rest].{key}.values[select2]")


# fit of cubic function to overdensity in l vs pm_b
# points chosen by eye because of large contamination
l_fit = [60, 50, 40, 30, 20, 10, 0, -10]
pm_b = [-18, -16, -13, -10, -5, 0, 5, 10]

popt2,pcov2 = curve_fit(cubic, l_fit, pm_b)
a_fit2, b_fit2, c_fit2, d_fit2 = popt2
print('pm_b x l fit')
print(popt2, np.sqrt(np.diag(pcov2)))

# new (broader) boundaries are set on distance and proper motion, no boundaries on l and b
select_dist2 = (df.phot_dist>2.5)& (df.phot_dist<5)
select_pm2 = (df.pm_l_apex>-40) & (df.pm_l_apex < -9) & (df.pm_b_apex>-35) & (df.pm_b_apex < 17)
select_mag2 = df.gmag_corrected<21

select_rest2 = Select_qual2 & select_dist2 & select_pm2 & select_mag2

# All stars within the mean + x standard deviation of the proper motion b from the fitted model
width =  np.mean(phleg_select2['pmb_uncertainty']) + 20*np.std(phleg_select2['pmb_uncertainty'])
select_fit = ((df.pmb > cubic(df.l_adj, a_fit2, b_fit2, c_fit2, d_fit2) - width) &    \
                    (df.pmb < cubic(df.l_adj, a_fit2, b_fit2, c_fit2, d_fit2) + width))

# all stars in the new selection that are within 10 deg of the stream track fitted
# earlier are selected
l_adj_arr2 = df[select_rest2 & select_fit].l_adj.values
b_arr2 = df[select_rest2 & select_fit].b.values
width = 10
select2_2 = np.where(((b_arr2 > quadratic(l_adj_arr2, a_fit, b_fit, c_fit) - width) &   \
             (b_arr2 < quadratic(l_adj_arr2, a_fit, b_fit, c_fit) + width)), True, False)
phleg_select3 = phleg_select_empty.copy()
for key in phleg_select3:
    phleg_select3[key] = eval(f"df[select_rest2 & select_fit].{key}.values[select2_2]")

# data without l x pm_b fit (only for plot)
l_adj_plot = df[select_rest2].l_adj.values
b_plot= df[select_rest2].b.values
select2_plot = np.where(((b_plot > quadratic(l_adj_plot, a_fit, b_fit, c_fit) - width) & \
      (b_plot < quadratic(l_adj_plot, a_fit, b_fit, c_fit) + width)), True, False)

phleg_plot = phleg_select_empty.copy()
for key in phleg_plot:
    phleg_plot[key] = eval(f"df[select_rest2].{key}.values[select2_plot]")

popt3, pcov3 = curve_fit(quadratic, phleg_select3["l_adj"], phleg_select3["b"])
a_fit3, b_fit3, c_fit3 = popt3


# improved fit to stream track, points still chosen by eye due to large contamination
l_orbit =  [60, 50, 40, 30, 20, 10, 0, -10, -20]
b_orbit = [-33, -39, -42, -43, -44, -43, -40, -36, -31]
popt4,pcov4 = curve_fit(quadratic, l_orbit, b_orbit)
a_fit4, b_fit4, c_fit4 = popt4

# all stars with in 6 degrees of the stream track are selected
width = 6
select4 = np.where(((b_arr2 > quadratic(l_adj_arr2, a_fit4, b_fit4, c_fit4) - width) &
             (b_arr2 < quadratic(l_adj_arr2, a_fit4, b_fit4, c_fit4) + width)), True, False)
phleg_select4 = phleg_select_empty.copy()
for key in phleg_select4:
    phleg_select4[key] = eval(f"df[select_rest2 & select_fit].{key}.values[select4]")

```

```
print('orbit fit 2:')
print(popt4, np.sqrt(np.diag(pcov4)))

# fit to overdensity in l vs. pm_l
# points picked by eye
l_fit2 = [60, 50, 40, 30, 20, 10, 0, -10]
pml = [-21, -28, -33, -36, -38, -38, -35, -31]

popt5,pcov5 = curve_fit(quadratic, l_fit2, pml)
a_fit5, b_fit5, c_fit5 = popt5
print('pm_l_x_l_fit')
print(popt5, np.sqrt(np.diag(pcov5)))

# All stars within the mean + x standard deviation of the proper motion l from the fitted model
width = np.mean(phleg_select4['pml_uncertainty']) + 7*np.std(phleg_select4['pml_uncertainty'])
select5 = np.where(((phleg_select4['pml'] > quadratic(phleg_select4['l_adj'], \
                  a_fit5, b_fit5, c_fit5) - width) & (phleg_select4['pml'] < \
          quadratic(phleg_select4['l_adj'], a_fit5, b_fit5, c_fit5) + width)), True, False)
phleg_select5 = phleg_select_empty.copy()
for key in phleg_select5:
    phleg_select5[key] = phleg_select4[key][select5]


# improved orbit fit

# Enough contamination is removed to not have the fit be based on hand picked points
# instead the data is put into bins of  l  = 10 deg, and the mean of b in each bin
# is taken to correspond to the centre of each bin
b_mean2 = np.array([])
i = 0
Sum = 0
bin_size = 10
bins_low2 = np.arange(region_sky[0][0], region_sky[0][1], bin_size)  #lower limits bins
bins = np.empty((len(bins_low2), len(phleg_select2)), dtype= object)
for l_low in bins_low2:
    select = np.where((phleg_select5["l_adj"] > l_low) & (phleg_select5["l_adj"]
                                          <= l_low + bin_size), True, False)

    j = 0
    for key in phleg_select5:
        bins[i][j] = phleg_select5[key][select]
        j += 1

    b_mean2 = np.append(b_mean2, np.median(bins[i][1]))
    i += 1

popt6,pcov6 = curve_fit(quadratic, bins_low2[1:7]+bin_size/2, b_mean2[1:7])
a_fit6, b_fit6, c_fit6 = popt6

print('orbit fit 3:')
print(popt6, np.sqrt(np.diag(pcov6)))

# All stars within 6 degrees of the newly fitted stream track are selected
width = 6
select6 = np.where(((phleg_select5['b'] > quadratic(phleg_select5['l_adj'],      \
                    a_fit6, b_fit6, c_fit6) - width) & (phleg_select5['b'] <   \
      quadratic(phleg_select5['l_adj'], a_fit6, b_fit6, c_fit6) + width)), True, False)
phleg_select6 = phleg_select_empty.copy()
for key in phleg_select6:
    phleg_select6[key] = phleg_select5[key][select6]


l_plot = np.linspace(-30,80,100)

# fit to overdensity in l vs. pm_ra
# Points on the overdensity are determined by taking the pm_ra value corresponding
# to the peak in the histogram for bins of  l  = 10 deg.

bin_size = 10
```

```
bins_low2 = np.arange(region_sky[0][0], region_sky[0][1], bin_size)  #lower limits bins
bins = np.empty((len(bins_low2), len(phleg_select2)), dtype= object)
pmra_mean2 = np.array([])
pmra_err_mean = np.array([])
pmra_err_std = np.array([])

#fit l vs. pmra
select7 = np.where(phleg_select6["l_adj"]>1000, False, False)
for l_low in bins_low2:
    select = np.where((phleg_select6["l_adj"] > l_low) & (phleg_select6["l_adj"]
                                            <= l_low + bin_size), True, False)

    pmra_counts, edges = np.histogram(phleg_select6['pmra'][select], bins=20)
    Max = edges[:-1][pmra_counts == np.max(pmra_counts)]+ (edges[1]-edges[0])/2
    pmra_mean2 = np.append(pmra_mean2, np.mean(Max))
    pmra_err_mean = np.append(pmra_err_mean, np.mean(phleg_select6['pmra_error'][select]))
    pmra_err_std = np.append(pmra_err_std, np.std(phleg_select6['pmra_error'][select]))

popt7,pcov7 = curve_fit(quadratic, bins_low2[2:]+bin_size/2, pmra_mean2[2:])
a_fit7, b_fit7, c_fit7 = popt7
print('fit l vs pmra')
print(popt7, np.sqrt(np.diag(pcov7)))

i=0
for l_low in bins_low2:
    # All stars within the mean + x standard deviation of the proper motion RA from the fitted model
    width = pmra_err_mean[i] + 3*pmra_err_std[i]
    select_2 =  np.where(((phleg_select6['pmra'] > quadratic(phleg_select6['l_adj'],    \
                           a_fit7, b_fit7, c_fit7) - width) & (phleg_select6['pmra'] < \
        quadratic(phleg_select6['l_adj'], a_fit7, b_fit7, c_fit7) + width)), True, False)
    select7 = select7 | select_2
    i += 1

phleg_select7 = phleg_select_empty.copy()
for key in phleg_select7:
    phleg_select7[key] = phleg_select6[key][select7]

# improved orbit fit, same method as before
b_mean3 =np.array([])
for l_low in bins_low2:
    select = np.where((phleg_select7["l_adj"] > l_low) & (phleg_select7["l_adj"]
                                              <= l_low + bin_size), True, False)

    b_mean3 = np.append(b_mean3, np.median(phleg_select7['b'][select]))

popt8,pcov8 = curve_fit(quadratic, bins_low2[1:7]+bin_size/2, b_mean3[1:7])
a_fit8, b_fit8, c_fit8 = popt8
print('orbit fit 4')
print(popt8, np.sqrt(np.diag(pcov8)))

width = 4
select8 = np.where(((phleg_select7['b'] > quadratic(phleg_select7['l_adj'],      \
                      a_fit8, b_fit8, c_fit8) - width) & (phleg_select7['b'] <   \
    quadratic(phleg_select7['l_adj'], a_fit8, b_fit8, c_fit8) + width)), True, False)
phleg_select8 =phleg_select_empty.copy()
for key in phleg_select8:
    phleg_select8[key] = phleg_select7[key][select8]

# fit l vs pmdec, final selection
# This time all data points are used to make the fit.
popt9,pcov9 = curve_fit(quadratic, phleg_select8['l_adj'], phleg_select8['pmdec'])
a_fit9, b_fit9, c_fit9 = popt9
print('fit l vs pmdec')
print(popt9, np.sqrt(np.diag(pcov9)))

# All stars within the mean + x standard deviation of the proper motion DEC from the fitted model
width = np.mean(phleg_select8['pmdec_error']) + 7*np.std(phleg_select8['pmdec_error'])
select9 =  np.where(((phleg_select8['pmdec'] > quadratic(phleg_select8['l_adj'],    \
                      a_fit9, b_fit9, c_fit9) - width) & (phleg_select8['pmdec'] <   \
```

```
344             quadratic(phleg_select8['l_adj'], a_fit9, b_fit9, c_fit9) + width)), True, False)
345 phleg_select9 = phleg_select_empty.copy()
346 for key in phleg_select9:
347     phleg_select9[key] = phleg_select8[key][select9]
348
349
350 # number of stars in final selection
351 print(len(phleg_select9['b']))
352
353 # final density map with gaussian smoothing
354 H, xedges, yedges = np.histogram2d(phleg_select9["l_adj"], phleg_select9["b"], \
355                                    bins=[int(110*2),int(50*2)], range=[[-30,80],[-60,-10]])
356 kernel = Gaussian2DKernel(x_stddev=1.5)  #gaussian smoothing
357 counts_smooth = convolve(H, kernel)
```